

ゲノム情報解析基礎

～ バイオインフォマティクス基礎知識とRのイントロ ～

¹大学院農学生命科学研究科
アグリバイオインフォマティクス教育研究プログラム

²微生物科学イノベーション連携研究機構

門田幸二(かどた こうじ)

kadota@iu.a.u-tokyo.ac.jp

<http://www.iu.a.u-tokyo.ac.jp/~kadota/>

講義予定

- 04月16日月曜日(17:15-20:30)
 - 嶋田透:ゲノムからの遺伝子予測
 - 門田幸二:バイオインフォマティクス基礎知識、Rのイントロダクション
- 04月23日月曜日(17:15-20:30)
 - 門田幸二:Rで塩基配列解析1、multi-FASTAファイルの各種解析
- 05月07日月曜日(17:15-20:30)
 - 嶋田透:ゲノムアノテーション、遺伝子の機能推定、RNA-seqなどによる発現解析、比較ゲノム解析
 - 門田幸二:Rで塩基配列解析2、Rパッケージ、k-mer解析の基礎
- 05月14日月曜日(17:15-19:00頃)
 - 勝間進:非コードRNA、小分子RNA、エピジェネティクス
 - 講義後、小テスト

各講義科目へのアクセス

①教育プログラム、②各講義のページ、③「ゲノム情報解析基礎」の場合



バイオインフォ関連情報

- ①「ゲノム情報解析基礎」のページ。
- ②前半はこのページを使います

2. ゲノム情報解析基礎

①

授業の目標・概要

次世代シーケンサーの普及により、ゲノム情報を基盤とした膨大な塩基配列情報を自在に解析するスキルが要求される時代になっています。フリーソフトRを用いて、配列決定後の基礎情報取得など各種配列解析の基本スキル向上を目指した実習を含む講義を行います。また、ウェブツールなどを用いて遺伝子領域の予測やアノテーションなどゲノム情報を比較または解析するための手法について解説します。

担当教員

嶋田 透 (東大・農・生産・環境生物学専攻 / 教授)
勝間 進 (東大・農・生産・環境生物学専攻 / 准教授)
門田幸二 (東大・農・アグリバイオ / 准教授)

参考図書

坊農秀雄 著、生命科学データ解析、MEDSI、2017

お知らせ

講義では、Rの様々なパッケージを利用します。
持ち込み用PC利用希望者は [インストール | について](#) を参考にしてR本体および必要なパッケージ群を必ずインストールしておいてください。

講義日程 (平成30年度)

1. 平成30年04月16日 (PC使用)

講師：嶋田 透

講師：門田幸二

[バイオインフォマティクス基礎知識](#)

[講義資料PDF\(Win版 ; 完全版\)](#)

[講義資料PDF\(Mac版 ; Rの説明部分のみ\)](#)

2. 平成30年04月23日 (PC使用)

②

[ゲノム情報解析基礎](#) バイオインフォマティクス基礎知識 (2018年04月16日) [門田幸二](#)

バイオインフォマティクス系学会など

- ・ [ISCB\(International Society for Computational Biology\)](#)
- ・ [JSBI\(日本バイオインフォマティクス学会\)](#)
- ・ ...

バイオインフォマティクス系?よろず相談所

- ・ [SEQanswers: Li et al., Bioinformatics, 2012](#)
 - NGSに特化した相談所。
- ・ [Biostar: Parnell et al., PLoS Comput Biol., 2011](#)
 - バイオインフォ全般。
- ・ [ライフサイエンスQA](#)
 - バイオインフォ全般。過疎ってる印象。
- ・ [Bio Technical フォーラム](#)
 - 主に実験系だがときどきバイオインフォ系トピックも。。。活発な印象。

バイオインフォマティクス人材育成カリキュラム(次世代シーケンサ)

- ・ バイオサイエンスデータベースセンター(NBDC)運営委員会人材育成分科会で、2014年3月に策定された[NGS用カリキュラム](#)が存在。このカリキュラムをベースにして、平成26-29年度にハンズオン(ノートPCを用いた実習型)講義を実施。おそらくアグリバイオ本体に次ぐ受講人数規模。
- ・ [NGS速習コース講習会](#) (平成26年度)
 - [NGS用カリキュラム](#)に沿った内容を東大農で2014年9月1日~12日に実施
 - 講習会映像は[統合TV](#)と [Youtube](#)から公開
 - [報告書](#)

①ISCB。②今後のイベント情報が分かります

学会(国外)

ゲノム情報解析基礎 バイオインフォマティクス基礎知識(2018年04月16日) 門田幸二

バイオインフォマティクス系学会など

- [ISCB\(International Society for Computational Biology\)](#)
- [JSBI\(日本バイオインフォマティクス学会\)](#)
- ...



バイオインフォマティクス系?よろず相談所

- [SEQanswers: Li et al., Bioinformatics, 2012](#)
 - NGSに特化した相談所。
- [Biostar: Parnell et al., PLoS Comput Biol., 2011](#)
 - バイオインフォ全般。
- [ライフサイエンスQA](#)
 - バイオインフォ全般。過疎ってる印象。
- [Bio Technical フォーラム](#)
 - 主に実験系だがときどきバイオインフォ系トピックも。。。活発な

バイオインフォマティクス人材育成カリキュラム(次世代シーク)

- [バイオサイエンスデータベースセンター\(NBDC\)運営委員会人材育成](#)で、2014年3月に策定された[NGS用カリキュラム](#)が存在。このカリキュラムにして、平成26-29年度にハンズオン(ノートPCを用いた実習)実施。おそらくアグリバイオ本体に次ぐ受講人数規模。
- [NGS速習コース講習会\(平成26年度\)](#)
 - [NGS用カリキュラム](#)に沿った内容を東大農で2014年9月1日～
 - 講習会映像は[統合TV](#)と [Youtube](#)から公開
 - [報告書](#)



学会(国内)

①JSBi。②今後のイベント情報が分かります

ゲノム情報解析基礎 バイオインフォマティクス基礎知識(2018年04月16日) 門田幸二

バイオインフォマティクス系学会など

- ・ [ISCB\(International Society for Computational Biology\)](#)
- ・ [JSBi\(日本バイオインフォマティクス学会\)](#)
- ・ ...



バイオインフォマティクス系?!よろず相談所

- ・ [SEQanswers: Li et al., Bioinformatics, 2012](#)
 - NGSに特化した相談所。
- ・ [Biostar: Parnell et al., PLoS Comput Biol., 2011](#)
 - バイオインフォ全般。
- ・ [ライフサイエンスQA](#)
 - バイオインフォ全般。過疎ってる印象。
- ・ [Bio Technical フォーラム](#)
 - 主に実験系だがときどきバイオインフォ系トピックも。。。活発な

バイオインフォマティクス人材育成カリキュラム(次世代シーク)

- ・ バイオサイエンスデータベースセンター(NBDC)運営委員会人材育成で、2014年3月に策定された[NGS用カリキュラム](#)が存在。このカリキュラムにして、平成26-29年度にハンズオン(ノートPCを用いた実習)実施。おそらくアグリバイオ本体に次ぐ受講人数規模。
- ・ [NGS速習コース講習会](#)(平成26年度)
 - [NGS用カリキュラム](#)に沿った内容を東大農で2014年9月1日～実施
 - 講習会映像は[統合TV](#)と [Youtube](#)から公開
 - [報告書](#)

The screenshot shows the homepage of the Japanese Society for Bioinformatics (JSBi). At the top, there is a navigation bar with a search box, an 'English' link, and a menu with 'ホーム', 'お問い合わせ', 'FAQ', and 'サイトマップ'. Below the navigation bar, there are three main columns of text describing the society's goals: 'バイオインフォマティクスに関する学問分野の発展', 'バイオインフォマティクスに関する技術および関連事業の振興', and 'バイオインフォマティクス教育の基盤確立と充実に寄与'. A central message states: '日本バイオインフォマティクス学会は、これらを目的として設立されました。詳しくは学会案内をご覧ください。' Below this, there is a horizontal menu with 'お知らせ', 'イベント案内', '研究会・地域部会', 'バイオインフォマティクス技術者認定', '学会案内', and '入会・継続案内'. A red arrow points to the 'イベント案内' link. The main content area is divided into two columns: '公募のお知らせ(人事関連)' and '最新イベント情報'. The '最新イベント情報' section lists several events, including the 7th IIBMP2018, the 2018 Japanese Environmental Chemistry Society workshop, and the 2018 Japanese Society for Bioinformatics workshop. A red arrow points to the '最新イベント情報' section. At the bottom, there is a section for '学会へのご意見・ご要望'.

相談窓口(国外)

①英語でググると、このどちらかのサイトによくヒットします。②SEQanswersは主にNGS関係のQAサイト

ゲノム情報解析基礎 バイオインフォマティクス基礎知識(2018年04月16日) 門田幸二

バイオインフォマティクス系学会など

- ・ [ISCB\(International Society for Computational Biology\)](#)
- ・ [JSBI\(日本バイオインフォマティクス学会\)](#)
- ・ ...

② バイオインフォマティクス系?!よろず相談所

- ・ [SEQanswers: Li et al., Bioinformatics, 2012](#)
 - ・ NGSに特化した相談所。
- ・ [Biostar: Parnell et al., PLoS Comput Biol., 2011](#)
 - ・ バイオインフォ全般。

- ・ [ライフサイエンスQA](#)
 - ・ バイオインフォ全般。過疎ってる印象。

- ・ [Bio Technical フォーラム](#)
 - ・ 主に実験系だがときどきバイオインフォ系トピックも。。。活発

バイオインフォマティクス人材育成カリキュラム(次世代シーク)

- ・ バイオサイエンスデータベースセンター(NBDC)運営委員会人材育成で、2014年3月に策定された[NGS用カリキュラム](#)が存在。このカリキュラムにして、平成26-29年度にハンズオン(ノートPCを用いた実習)実施。おそらくアグリバイオ本体に次ぐ受講人数規模。
- ・ [NGS速習コース講習会](#)(平成26年度)
 - ・ [NGS用カリキュラム](#)に沿った内容を東大農で2014年9月1日～
 - ・ 講習会映像は[統合TV](#)と [Youtube](#)から公開
 - ・ [報告書](#)

Forum	Last Post	Threads	Posts
Introductions (8 Viewing) New here? Stop in and introduce yourself. Where you are, what you work on, etc.	Introduction by PaulHudson 03-09-2018 01:21 AM	2,017	4,029
General (11 Viewing) Any topic/question that does not fit into the subcategories below. If you're unsure of where to put something, ask in here!	Anilent 4200 TapeStation... by lmi 03-07-2018 06:53 PM	1,663	6,542
Core Facilities Dealing with customer samples, data, and the challenges that come with both?	Looking for extra BioAnalyzer... by Christine 02-06-2018 10:40 AM	116	539
Sequencing Technologies/Companies Platform specific questions, news, discussions			
454 Pyrosequencing (6 Viewing) Pyrosequencing in picotiter plates, custom arrays for enrichment/decomplexing. (Roche)	454 data analysis in Qiime1 by LMM 01-11-2018 06:30 AM	740	3,948
Complete Genomics (1 Viewing) Service-based whole human genome sequencing with probe ligation chemistry.	The title of this forum... by zhengang 12-09-2017 06:03 PM	45	150
Helicos / Direct Genomics True Single Molecule Sequencing, originally by Helicos, tech acquired by Direct Genomics	Nature Biotech:Direct... by DebbyPan 02-13-2016 05:21 PM	26	143
Illumina/Solexa (29 Viewing) Bridged amplification & clustering followed by sequencing by synthesis. (Genome Analyzer / HiSeq / MiSeq)	Flow cell by Kultamans2 03-11-2018 06:53 AM	2,636	15,328
Ion Torrent (2 Viewing) Integrated electronic detection of protons released upon base incorporation. (Life Technologies)	Ion Proton S5 Chips - 540 by rrosati 02-26-2018 05:52 AM	343	2,056
Oxford Nanopore (2 Viewing) Single molecule sequencing via a protein nanopore, presently on the MinION.	Favourite Software by ilisanni 02-05-2018 12:50 AM	62	441
Pacific Biosciences (2 Viewing) Single-molecule real-time observation of DNA polymerase using zero-mode	pbJelly by bin_d 03-05-2018 10:48 PM	328	2,044

①国内版もあるので、②などを有効利用してはいかが

相談窓口(国内)

ゲノム情報解析基礎 バイオインフォマティクス基礎知識(2018年04月16日) 門田幸二

バイオインフォマティクス系学会など

- ・ [ISCB\(International Society for Computational Biology\)](#)
- ・ [JSBI\(日本バイオインフォマティクス学会\)](#)
- ・ ...

バイオインフォマティクス系?よろず相談所

- ・ [SEQanswers: Li et al., Bioinformatics, 2012](#)
 - NGSに特化した相談所。
- ・ [Postar: Parnell et al., PLoS Comput Biol., 2011](#)
 - バイオインフォ全般。
- ② [ライフサイエンスQA](#)
 - バイオインフォ全般。過疎ってる印象。
- ① [Bio Technical フォーラム](#)
 - 主に実験系だがときどきバイオインフォ系トピックも。。。活発な印象。

バイオインフォマティクス人材育成カリキュラム(次世代シーケンサ)

- ・ バイオサイエンスデータベースセンター(NBDC)運営委員会人材育成分科会で、2014年3月に策定された[NGS用カリキュラム](#)が存在。このカリキュラムをベースにして、平成26-29年度にハンズオン(ノートPCを用いた実習型)講義を実施。おそらくアグリバイオ本体に次ぐ受講人数規模。
- ・ [NGS速習コース講習会\(平成26年度\)](#)
 - [NGS用カリキュラム](#)に沿った内容を東大農で2014年9月1日~12日に実施
 - 講習会映像は[統合TV](#)と [Youtube](#)から公開
 - [報告書](#)



NGS用カリキュラム

2014年3月に①NBDCによって策定された②NGS用カリキュラム。最低限必要とされる知識・技術を2週間程度で身につけることを想定した「速習」と「速習以外」に分かれている

バイオインフォマティクス人材育成カリキュラム(次世代シーケンサ)

・バイオサイエンスデータベースセンター(NBDC)運営委員会人材育成分科会で、2014年3月に策定された**NGS用カリキュラム**②をベースにして、平成26-29年度にハンズオン(ノー)①を実施。おそらくアグリバイオ本体に次ぐ受講人数規模。

- ・ **NGS速習コース講習会**(平成26年度)
 - ・ **NGS用カリキュラム**に沿った内容を東大農で2014年実施
 - ・ 講習会映像は**統合TV**と **Youtube**から公開
 - ・ **報告書**
- ・ **NGS/ハンズオン講習会**(平成27年度)
 - ・ 東大農で2015年07月22日～08月06日(A日程)および(B日程)に実施
 - ・ 講習会映像は**統合TV**と **Youtube**から公開
 - ・ **報告書**
- ・ **NGS/ハンズオン講習会**(平成28年度)
 - ・ 東大農で2016年07月19日～08月04日に実施
 - ・ 講習会映像は**統合TV**と **Youtube**から公開
 - ・ **報告書**
- ・ **NGS/ハンズオン講習会**(平成29年度)
 - ・ 東大農で2017年08月28日～09月01日に実施
 - ・ 講習会映像は**統合TV**と **Youtube**から公開
 - ・ **報告書**

バイオインフォマティクス人材育成カリキュラム(次世代シーケンサ)

本カリキュラムは、次世代シーケンサデータを扱うにあたり最低限必要とされる知識・技術を2週間程度で身につけることを想定した「速習」と、時間をかけて習得することを想定した「速習以外」に分かれています。

【速習】

大項目	日数	No.	項目	習得技術	レベル	形式
1. コンピュータリテラシーとサーバー設計	4日	1-1	OS、ハード構成	・コンピュータの基本的理解	初級	講義
		1-2	ネットワーク基礎	・インターネット、セキュリティの基本的理解	初級	講義
		1-3	UNIX I	・UNIXの基礎的理解 ・Linux導入	中級	実習
	2日	1-4	スクリプト言語	・Perl ・シェルスクリプト	中級	実習
2. 配列インフォマティクス	1日	2-1	配列解析基礎	・配列、ゲノムデータ記述のフォーマット、アラインメント(DP)、データベース検索(BLAST, BLAT)等の基礎的な配列比較解析の原理と実習	初級	実習
		2-2	バイオ系データベース概論	・基本的な各種バイオ系データベースの理解、統合DBの利用法	初級	実習
3. データ解析基礎	2日	3-1	R 基礎1	・R言語の基礎(インストールから利用まで)	初級	実習
		3-2	R 基礎2	・ファイルの読み込み、行列演算の基本	初級	実習
		3-3	R 各種パッケージ	・Rの各種パッケージのインストール法と代表的なパッケージの利用法	中級	実習
		3-4	R bioconductor I	・bioconductorの利用法	中級	実習
		3-5	R bioconductor II	・FASTA and FASTQ形式ファイルの読み込み、ファイル形式の変換(FASTQ → FASTA)、クオリティチェック、リード配列長分布、フィルタリングやトリミング、GC含量計算など	中級	実習
4. 次世代シーケンサ	2.5日	4-1	次世代シーケンサ基礎I	・原理の理解	初級	講義
		4-2	次世代シーケンサ基礎II	・応用分野とそれのための計測技術の理解(RNA-seq, ChIP-seq, がんゲノム, 個人ゲノム, 環境ゲノム, Hi-C)	初級	講義
	0.5日	4-3	次世代シーケンサ実習I	・ファイル形式、可視化、quality check、マッピング、アセンブル	初級	実習
	1.5日	4-4	次世代シーケンサ実習II	・代表的なパイプラインについての実習(多型解析(IGV)、RNA-seq、ChIP-seq、及び統合解析、定番のツールを利用)	初級	実習
5. ゲノム関連の倫理・法律	0.5日	5-1	ゲノム情報倫理概論	・ゲノム情報を扱う上で、プライバシー保護などの必要な倫理的問題、法的問題の国内外の状況を理解し、ゲノム情報を適切に利用できるようにする ・匿名化、暗号化、情報セキュリティ概要	初級	講義
6. 分子生命科学	0.5日	6-1	分子生命科学概論	・複製、転写、翻訳、代謝、シグナル伝達などの基礎知識	初級	講義
		6-2	オミクス概論	・ゲノム以外のオミクスデータの基礎知識	初級	講義
		6-3	遺伝/進化概論	・ゲノムデータを扱う上での遺伝学、進化学の基礎知識	初級	講義



NGS講習会

①NBDC主導で4年間にわたり開催されたNGS講習会。各年度の講習会内容は②から辿れる

バイオインフォマティクス人材育成カリキュラム(次世代シーケンサ)

- ・ バイオサイエンスデータベースセンター(NBDC)運営委員会人材育成分科会で、2014年3月に策定されたNGS用カリキュラムが存在。このカリキュラムをベースにして、平成26-29年度にハンズオン(ノートPCを用いた実習型)講義を実施。おそらくアグリ^①本体に次ぐ受講人数規模。
- ・ NGS速習コース講習会(平成26年度)
 - NGS用カリキュラムに沿った内容を東大農で2014年9月1日～12日に実施
 - 講習会映像は統合TVと Youtubeから公開
 - 報告書
- ・ NGS/ハンズオン講習会(平成27年度)
 - 東大農で2015年07月22日～08月06日(A日程)および08月26-28日(B日程)に実施
 - 講習会映像は統合TVと Youtubeから公開
 - 報告書
- ・ NGS/ハンズオン講習会(平成28年度)
 - 東大農で2016年07月19日～08月04日に実施
 - 講習会映像は統合TVと Youtubeから公開
 - 報告書
- ・ NGS/ハンズオン講習会(平成29年度)
 - 東大農で2017年08月28日～09月01日に実施
 - 講習会映像は統合TVと Youtubeから公開
 - 報告書

平成29年度NGS講習会

例えば①平成29年度の講習会の講義資料(や動画)は、②から見られます

バイオインフォマティクス人材育成カリキュラム(次世代シーケンサ)

- ・バイオサイエンスデータベースセンター(NBDC)運営委員会人材育成分科会で、2014年3月に策定されたNGS用カリキュラムが存在。このカリキュラムをベースにして、平成26-29年度にハンズオン(ノートPCを用いた実習型)講義を実施。おそらくアグリバイオ本体に次ぐ受講人数規模。
- ・NGS速習コース講習会(平成26年度)
 - NGS用カリキュラムに沿った内容を東大農で2014年実施
 - 講習会映像は統合TVとYoutubeから公開
 - 報告書
- ・NGSハンズオン講習会(平成27年度)
 - 東大農で2015年07月22日～08月06日(A日程)おこなう(実施)
 - 講習会映像は統合TVとYoutubeから公開
 - 報告書
- ・NGSハンズオン講習会(平成28年度)
 - 東大農で2016年07月19日～08月04日に実施
 - 講習会映像は統合TVとYoutubeから公開
 - 報告書
- ・NGSハンズオン講習会(平成29年度)
 - 東大農で2017年08月28日～09月01日に実施
 - 講習会映像は統合TVとYoutubeから公開
 - 報告書

The screenshot shows the NBDC website interface. At the top, there is a navigation bar with links for Home, NBDCについて, 研究開発, 公募情報, 採用情報, イベント, 人材支援, アクセス, and リンク. The main content area is titled "平成29年度NGSハンズオン講習会" and includes a "お知らせ" (Notice) section with a list of dates and events. Below this is a "本ページの目次" (Table of Contents) section with a table of links: 開催概要, 受講申し込み方法, カリキュラム・講義資料・動画, and お問い合わせ先. The "開催概要" section is partially visible at the bottom.

開催概要
○概要 : Linux 上で次世代シーケンサー(NGS)データを自在に解析するための実習を行います。今年度は、ゲノムアセンブリ後の各種解析手法について日本乳酸菌学会誌の連載「次世代シーケンサーデータの解析手法」第7回の一部及び第8回を中心とした講義と、メタゲノム解析、Hi-C解析についての講義を実施します。受講者には、講師指定の事前予習を行っていただきます。 なお、NGSの概論的な講義(リード、ペアエンドなどの基本的な用語やNGS原理の説明など)は省略します。過去の年度の講義資料や参考図書など参照の上、基礎知識を取得しておくことをお勧めします。

ゲノム情報解析 ≡ NGS解析

①「ゲノム情報解析基礎」
で教えられる内容は、②
NGS解析全体のごく一部

バイオインフォマティクス人材育成カリキュラム(次世代シーケンサ)

- ・ バイオサイエンスデータベースセンター(NBDC)運営委員会人材育成分科会で、2014年3月に策定されたNGS用カリキュラムが存在。このカリキュラムをベースにして、平成26-29年度にハンズオン(ノートPCを用いた実習型)講義を実施。おそらくアグリバイオ本体に次ぐ受講人数規模。
- ・ **NGS速習コース講習会**(平成26年度)
 - **NGS用カリキュラム**に沿った内容を東大農で2014年9月1日～12日に実施
 - 講習会映像は**統合TV**と **Youtube**から公開
 - **報告書**
- ・ **NGS/ハンズオン講習会**(平成27年度)
 - 東大農で2015年07月22日～08月06日(A日程)および08月26-28日(B日程)に実施
 - 講習会映像は**統合TV**と **Youtube**から公開
 - **報告書**
- ・ **NGS/ハンズオン講習会**(平成28年度)
 - 東大農で2016年07月19日～08月04日に実施
 - 講習会映像は**統合TV**と **Youtube**から公開
 - **報告書**
- ・ **NGS/ハンズオン講習会**(平成29年度)
 - 東大農で2017年08月28日～09月01日に実施
 - 講習会映像は**統合TV**と **Youtube**から公開
 - **報告書**

アグリバイオの教育プログラム



ゲノム情報解析

①この科目では、フリーソフトRで塩基配列解析を行う基本スキルの伝授のみ。②「バイオスタティスティクス基礎論」でもRを使いますが、統計解析用と塩基配列解析用は方向性や使用感が随分異なる。アグリバイオは主にRで教えるという設立当初からの全体方針(歴史的背景)があります

バイオインフォマティクス人材育成カリキュラム(次世代)

- ・バイオサイエンスデータベースセンター(NBDC)運営委員で、2014年3月に策定されたNGS用カリキュラムが存在。このカリキュラムをベースにして、平成26-29年度に(ハンズオン(ノートPCを用いた実習型)講義を実施。おそらくアグリバイオ本体に次ぐ受講人数規模。
- ・NGS速習コース講習会(平成26年度)
 - NGS用カリキュラムに沿った内容を東大農で2014年9月1日~12日に実施
 - 講習会映像は統合TVと Youtubeから公開
 - 報告書
- ・NGSハンズオン講習会(平成27年度)
 - 東大農で2015年07月22日~08月06日(A日程)および08月26-28日(B日程)に実施
 - 講習会映像は統合TVと Youtubeから公開
 - 報告書
- ・NGSハンズオン講習会(平成28年度)
 - 東大農で2016年07月19日~08月04日に実施
 - 講習会映像は統合TVと Youtubeから公開
 - 報告書
- ・NGSハンズオン講習会(平成29年度)
 - 東大農で2017年08月28日~09月01日に実施
 - 講習会映像は統合TVと Youtubeから公開
 - 報告書

アグリバイオの教育プログラム



講習会関連

①(おそらくこれ以外にも)各自の事情や感性に合った講習会があると思います。教え方はひとそれぞれなので色々出られてみてはいかがでしょうか

講習会関連

- ・ [イリミナ株式会社のiSchool](#)
 - NGS機器のシェアNo.1企業提供の講習会やウェビナー。クラウド解析環境 [BaseSpace](#) なども提供している。
- ・ [基礎生物学研究所のトレーニングコース](#)
 - おそらく受講生あたりのスタッフ数をもっとも充実しているコース。
- ・ DNA Data Bank of Japan (DDBJ)センターが提供する各種講習会
 - [DDBJing講習会](#)、[D-STEP講習会](#) などいくつかあります。遺伝研スパコンの使い方を知りたい場合はこちら。



自習用教材

- ・ [\(Rで\)塩基配列解析](#)
 - NGS系のそれっぽいキーワードで検索すると大抵上位に出現。RのインストールからLinux環境でのNGS解析まで幅広く解説。
 - [日本乳酸菌学会誌](#)上で2014年から連載中のNGS記事も [こちら](#) で提供。
- ・ [統合TV](#)
 - 有用なデータベースやウェブツールの活用法を動画で紹介してくれます。
- ・ [Biopyrus](#)
 - ここもNGS系のそれっぽいキーワードで検索すると大抵上位に出現。Linux、Perlなどの各種インストール系や用語説明などが豊富。

自習用教材

講習会関連

- ・ [イルミナ株式会社](#)の*iSchool*
 - NGS機器のシェアNo.1企業環境 [BaseSpace](#) なども提供して
- ・ [基礎生物学研究所](#)の*トレーニング*
 - おそらく受講生あたりのスタ
- ・ DNA Data Bank of Japan ([DDBJ](#))
 - [DDBJing 講習会](#)、[D-STEP](#) などの使い方を知りたい場合は

(Rで)塩基配列解析

(last modified 2018/03/12, since 2010)

このウェブページのR関連部分は、[インストール |](#) についての推奨手順 ([Windows2018.03.12版](#)と[Macintosh2015.04.03版](#))に従ってフリーソフトRと必要なパッケージをインストール済みであるという前提で記述しています。初心者の方は [基本的な利用法](#) ([Windows2015.04.03版](#)と [Macintosh2015.04.03版](#))で自習してください。本ウェブページを体系的にまとめた [書籍](#) もあります。(2015/04/03)

What's new?

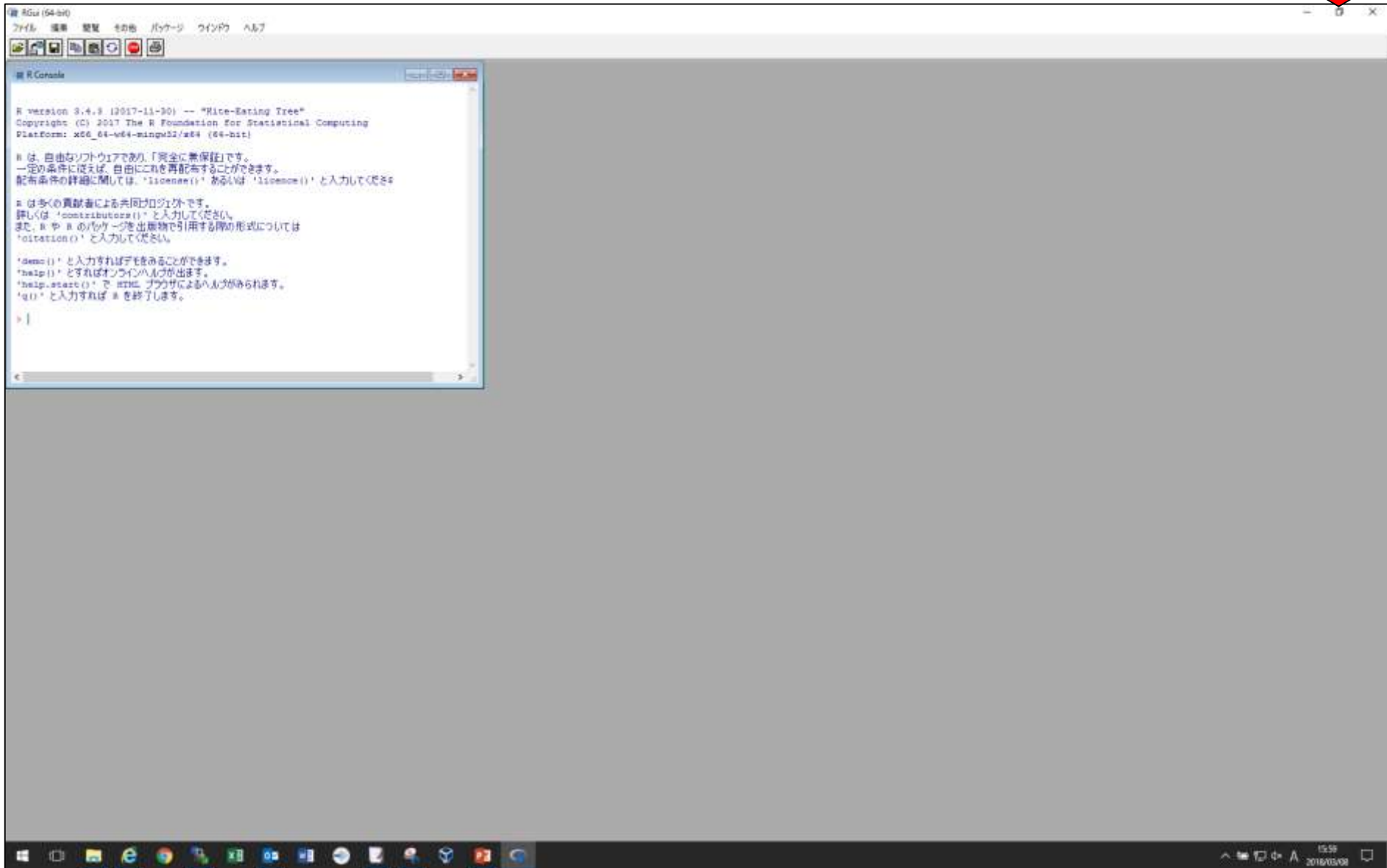
- ・ [アグリバイオインフォマティクス教育研究プログラム](#)では、平成30年度もバイオインフォ関連講義を行います。例年東大以外の企業の方、研究員、大学院生が2-3割程度受講しております。受講ガイダンスは、平成30年4月4日17:15より東大農学部2号館2階化学第一講義室で開催します。(2018/03/08) **NEW**
- ・ Silhouetteスコアの新たな使い道提唱論文([Zhao et al., Biol. Proc. Online, 2018](#))の使い方を「解析 | 一般 | [Silhouette scores](#)(シルエットスコア)」に示しました。(2018/03/01) **NEW**
- ・ Silhouetteスコアの新たな使い道提唱論文([Zhao et al., Biol. Proc. Online, 2018](#))がpublishされました。(2018/03/01) **NEW**
- ・ 「平成29年度[NGSハンズオン講習会](#)」の [動画](#) が公開されています。(2018/03/01) **NEW**

- ・ [門田からメール返信をもらえない場合は](#) (last modified 2016/08/23)
- ・ [はじめに](#) (last modified 2015/03/31)
- ・ 参考資料 | [書籍、学会誌](#) (last modified 2017/11/13)
- ・ 参考資料 | [講習会、講義、講演資料](#) (last modified 2017/09/07)
- ・ [過去のお知らせ](#) (last modified 2018/03/01) **NEW**
- ・ [インストール |](#) について (last modified 2018/03/12) **NEW**
- ・ インストール | R本体 | 最新版 | [Win用](#) (last modified 2015/03/22) 推奨
- ・ インストール | R本体 | 最新版 | [Mac用](#) (last modified 2015/04/22) 推奨
- ・ インストール | R本体 | 過去版 | [Win用](#) (last modified 2015/03/22)
- ・ インストール | R本体 | 過去版 | [Mac用](#) (last modified 2015/03/22)

自習用教材

- ・ [\(Rで\)塩基配列解析](#)
 - NGS系のそれっ(まい)キーワードからLinux環境での
 - [日本乳酸菌学会誌](#) 上で2014
- ・ [統合TV](#)
 - 有用なデータベースやウェブ
- ・ [Biopapyrus](#)
 - ここもNGS系のそれっ(まい)キ
 - Linux、Perlなどの各種インス

Rの起動



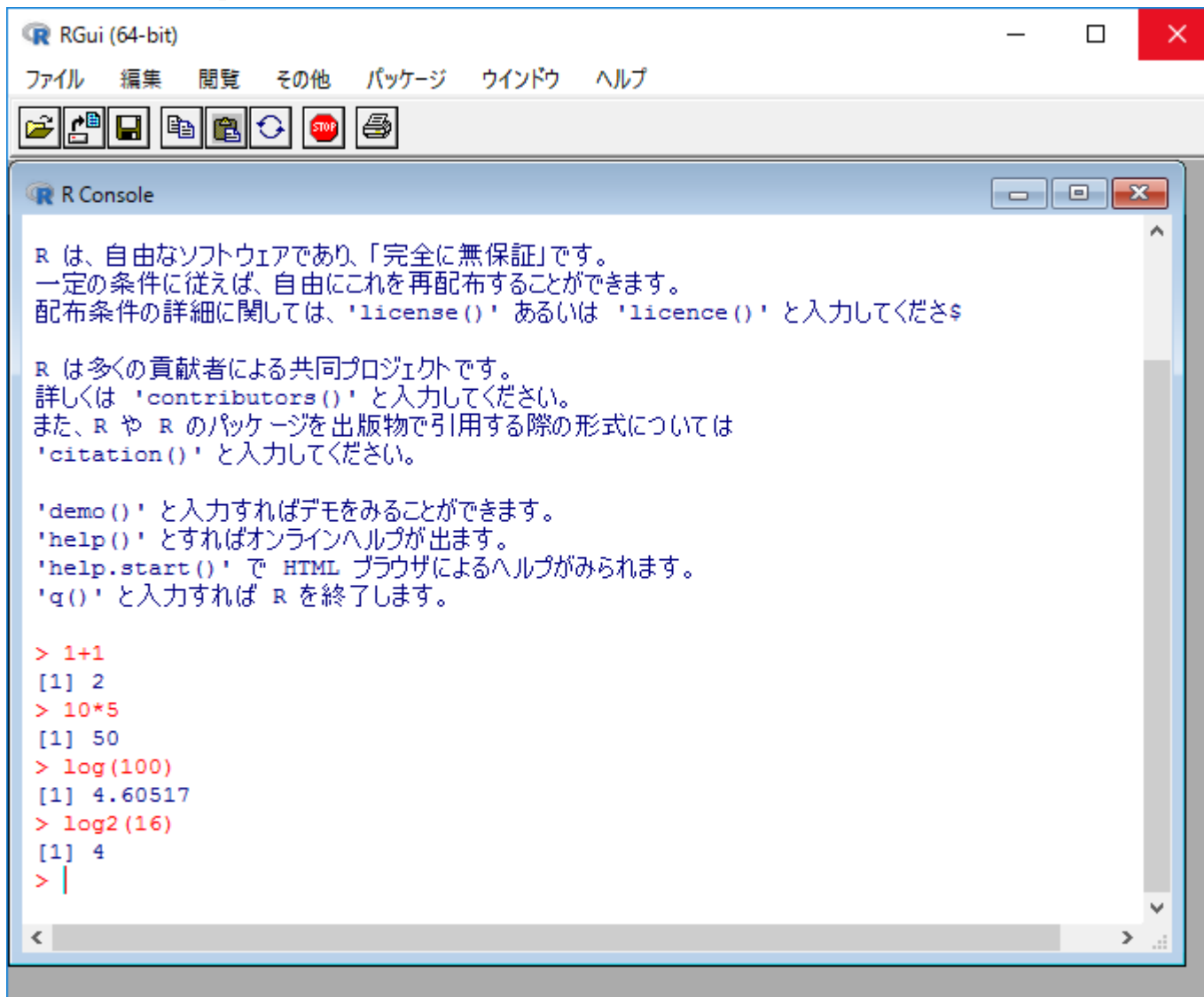
Rの起動



①赤枠で囲まれた部分が「Rコンソール画面」

```
R version 3.4.3 (2017-11-30) -- "Kite-Eating Tree"  
Copyright (C) 2017 The R Foundation for Statistical Computing  
Platform: x86_64-w64-mingw32/x64 (64-bit)  
  
R は、自由なソフトウェアであり、「完全に無保証」です。  
一定の条件に従えば、自由にこれを再配布することができます。  
配布条件の詳細に関しては、'license()' あるいは 'licence()' と入力してくださ$  
  
R は多くの貢献者による共同プロジェクトです。  
詳しくは 'contributors()' と入力してください。  
また、R や R のパッケージを出版物で引用する際の形式については  
'citation()' と入力してください。  
  
'demo()' と入力すればデモをみることができます。  
'help()' とすればオンラインヘルプが出ます。  
'help.start()' で HTML ブラウザによるヘルプがみられます。  
'q()' と入力すれば R を終了します。  
  
> |
```

基本的な利用法



RGui (64-bit)

ファイル 編集 閲覧 その他 パッケージ ウィンドウ ヘルプ

R Console

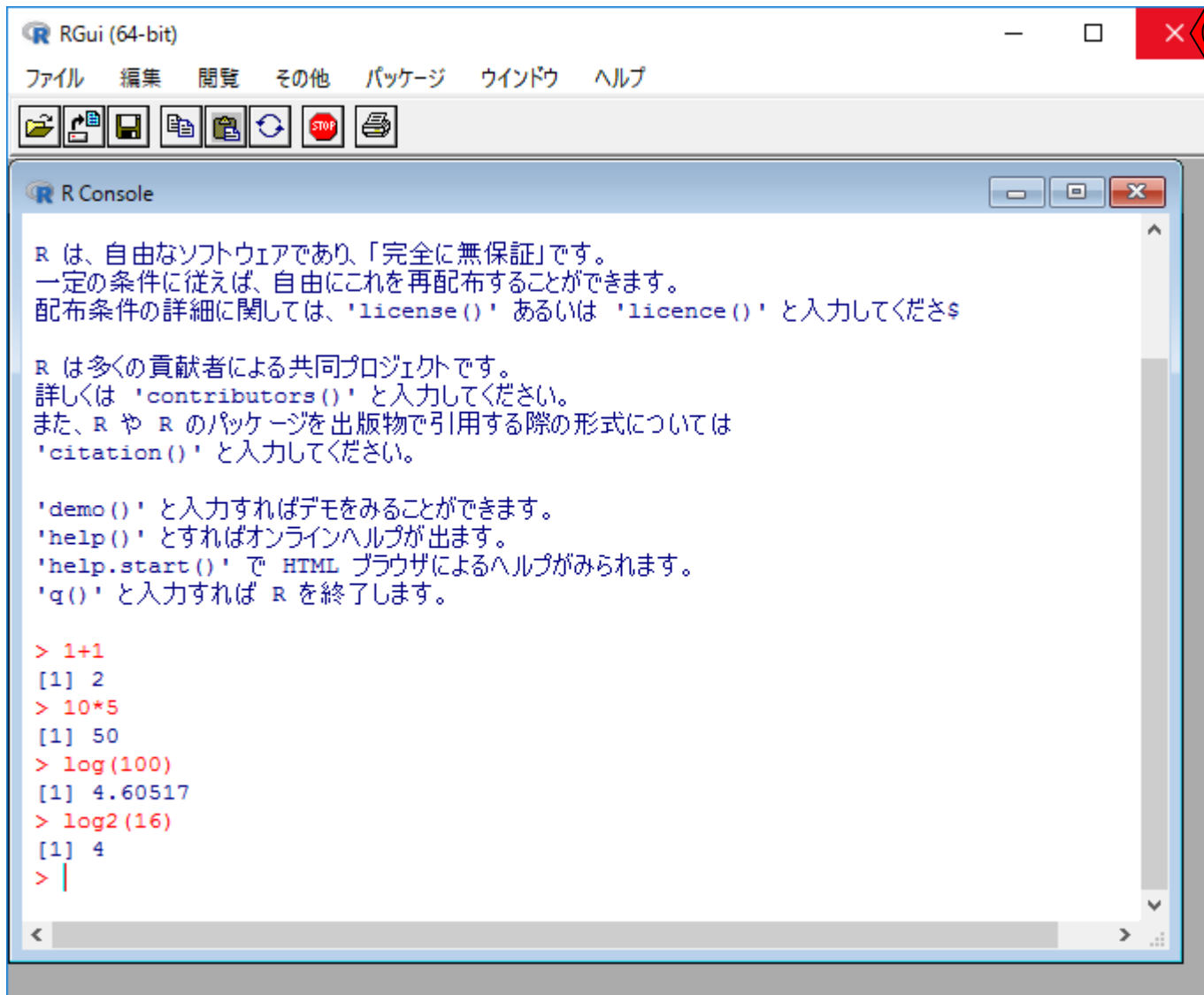
R は、自由なソフトウェアであり、「完全に無保証」です。
一定の条件に従えば、自由にこれを再配布することができます。
配布条件の詳細に関しては、`'license()'` あるいは `'licence()'` と入力してください\$

R は多くの貢献者による共同プロジェクトです。
詳しくは `'contributors()'` と入力してください。
また、R や R のパッケージを出版物で引用する際の形式については
`'citation()'` と入力してください。

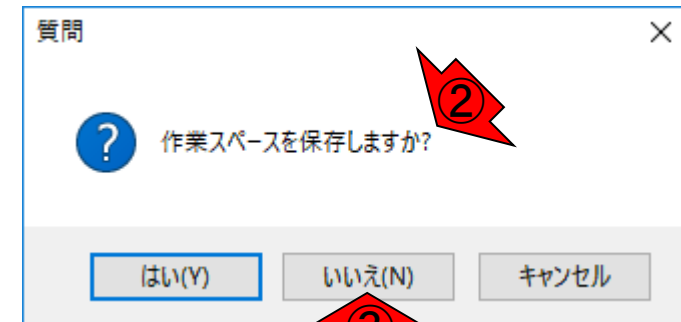
`'demo()'` と入力すればデモをみることができます。
`'help()'` とすればオンラインヘルプが出ます。
`'help.start()'` で HTML ブラウザによるヘルプがみられます。
`'q()'` と入力すれば R を終了します。

```
> 1+1
[1] 2
> 10*5
[1] 50
> log(100)
[1] 4.60517
> log2(16)
[1] 4
> |
```

Rの終了



通常のソフトウェアと同様、①
右上の×ボタンを押せばよい。
②「作業スペースを保存しま
すか?」というダイアログが出
るが、最初のうちは③いいえ
でよい。「はい」を押してし
まっても.Rdataと.Rhistoryとい
う2つのファイルが作成される
だけなので特に問題はない



①塩基配列を入力として、その翻訳されたアミノ酸配列を取得することができます

解析基礎1: 翻訳配列取得

- ・ [イントロ](#) | [一般](#) | [任意の位置の塩基を置換](#) (last modified 2013/09/12)
- ・ [イントロ](#) | [一般](#) | [指定した範囲の配列を取得](#) (last modified 2015/04/06)
- ・ [イントロ](#) | [一般](#) | [指定したID\(染色体やdescription\)の配列を取得](#) (last modified 2014/03/10)
- ・ [イントロ](#) | [一般](#) | [翻訳配列\(translate\)を取得\(基礎\)](#) | [Biostrings](#) **①** (last modified 2015/09/12)
- ・ [イントロ](#) | [一般](#) | [翻訳配列\(translate\)を取得\(応用\)](#) | [seqinr\(Chambers 2005\)](#) (last modified 2015/03/09)
- ・ [イントロ](#) | [一般](#) | [相補鎖\(complement\)を取得](#) (last modified 2013/06/14)
- ・ [イントロ](#) | [一般](#) | [逆相補鎖\(reverse complement\)を取得](#) (last modified 2013/06/14)
- ・ [イントロ](#) | [一般](#) | [逆鎖\(reverse\)を取得](#) (last modified 2013/06/14)
- ・ [イントロ](#) | [一般](#) | [k-mer解析](#) | [k=1\(塩基ごとの出現頻度\)](#)
- ・ [イントロ](#) | [一般](#) | [k-mer解析](#) | [k=2\(2連続塩基の出現頻度\)](#)
- ・ [イントロ](#) | [一般](#) | [k-mer解析](#) | [k=3\(3連続塩基の出現頻度\)](#)
- ・ [イントロ](#) | [一般](#) | [k-mer解析](#) | [k=n\(n連続塩基の出現頻度\)](#)
- ・ (削除予定) [イントロ](#) | [一般](#) | [2連続塩基の出現頻度情報](#)
- ・ (削除予定) [イントロ](#) | [一般](#) | [3連続塩基の出現頻度情報](#)
- ・ (削除予定) [イントロ](#) | [一般](#) | [任意の長さの連続塩基の出現頻度情報](#)
- ・ [イントロ](#) | [一般](#) | [Tips](#) | [任意の拡張子でファイルを保存](#)
- ・ [イントロ](#) | [一般](#) | [Tips](#) | [拡張子は同じで任意の文字を追加](#)

イントロ | 一般 | 翻訳配列(translate)を取得(基礎) | Biostrings **NEW**

[Biostrings](#)パッケージを用いて塩基配列を読み込んでアミノ酸配列に翻訳するやり方を示します。翻訳のための**遺伝コード(genetic code)**は、Standard Genetic Codeだそうです。もちろん生物種?!!によって多少違い(variants)があるようで、"Standard", "SGC0", "Vertebrate Mitochondrial", "SGC1"などいろいろ選べるようです。「ファイル」→「ディレクトリの変更」で解析したいファイルを置いてあるディレクトリに移動し以下をコピー。

1. FASTA形式ファイル(sample1.fasta)の場合:

multi-FASTAではないsingle-FASTA形式ファイルです。

```
in_f <- "sample1.fasta" #入力ファイル名を指定してin_fに格納
out_f <- "hoge1.fasta" #出力ファイル名を指定してout_fに格納

#必要なパッケージをロード
library(Biostrings) #パッケージの読み込み

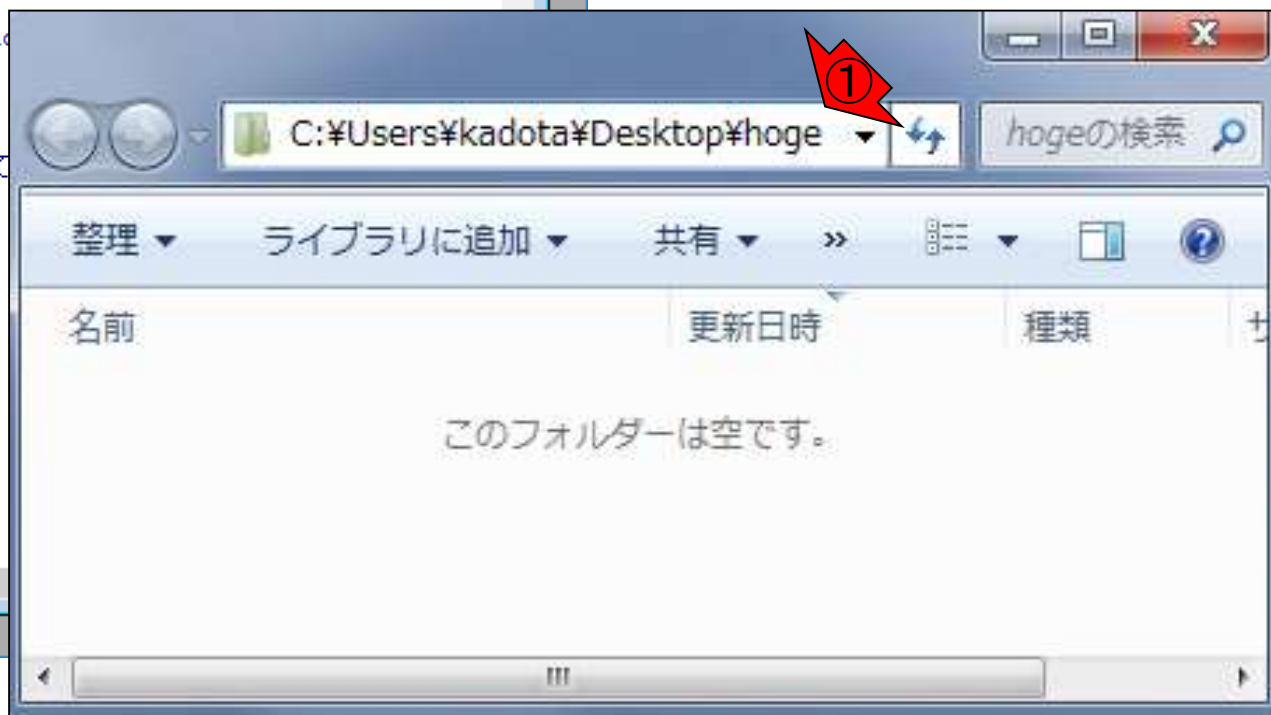
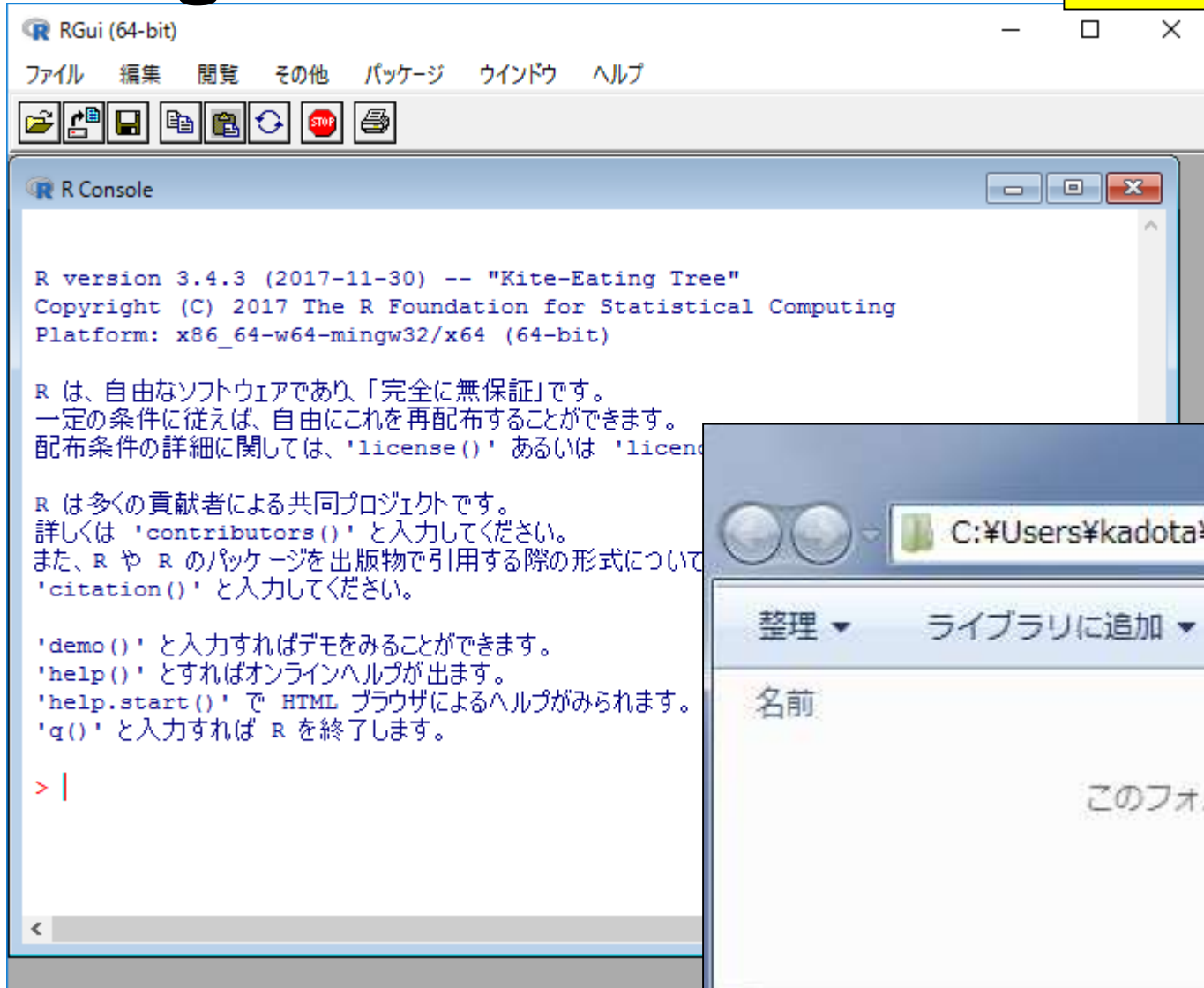
#入力ファイルの読み込み
fasta <- readDNASTringSet(in_f, format="fasta") #in_fで指定したファイルの読み込み
fasta #確認してるだけです

#本番
fasta <- translate(fasta) #アミノ酸配列に翻訳した結果をfastaに格納
fasta #確認してるだけです

#ファイルに保存
writeXStringSet(fasta, file=out_f, format="fasta", width=50) #fastaの中身を指定したフ
```

hogeフォルダの作成

デスクトップにあるhogeフォルダ中のファイルを解析するやり方として説明します。
① デスクトップ上にhogeフォルダを作成



ファイルの保存

①解析したいsample1.fastaのファイル名部分で
右クリックして②対象をファイルに保存。③デスク
トップ上に作成した④hogeフォルダに⑤保存

I. FASTA形式ファイル(sample1.fasta)の場合:

```
multi-FASTAではないsingle FASTAの場合  
in_f <- "sample1.fasta"  
out_f <- "hoge1.fasta"  
  
#必要なパッケージをロード  
library(Biostrings)  
  
#入力ファイルの読み込み  
fasta <- readDNASTringSet(  
  fasta  
  
#本番  
fasta <- translate(fasta)  
fasta  
  
#ファイルに保存  
writeXStringSet(fasta, file)
```

- 開く(O)
- 新しいタブで開く(W)
- 新しいウィンドウで開く(N)
- 対象をファイルに保存 **②**
- 対象を印刷(P)
- 切り取り
- コピー(C)
- ショートカットのコピー(T)
- 貼り付け(P)
- Bing で翻訳
- 電子メール (Windows Live)
- すべてのアクセラレータ
- 要素の検査(L)

名前を付けて保存

デスクトップ

整理 ▾ 新しいフォルダー

- お気に入り
 - SkyDrive
 - ダウンロード
 - デスクトップ **③**
 - 最近表示した場所
- ライブラリ
 - ドキュメント
 - ピクチャ
 - ビデオ
 - ミュージック

ライブラリ

- ライブラリ
- システム フォルダー
 - kadota
 - コンピュータ
 - ネットワーク
- hoge **④**
ファイル フォルダー

ファイル名(N): sample1.fasta

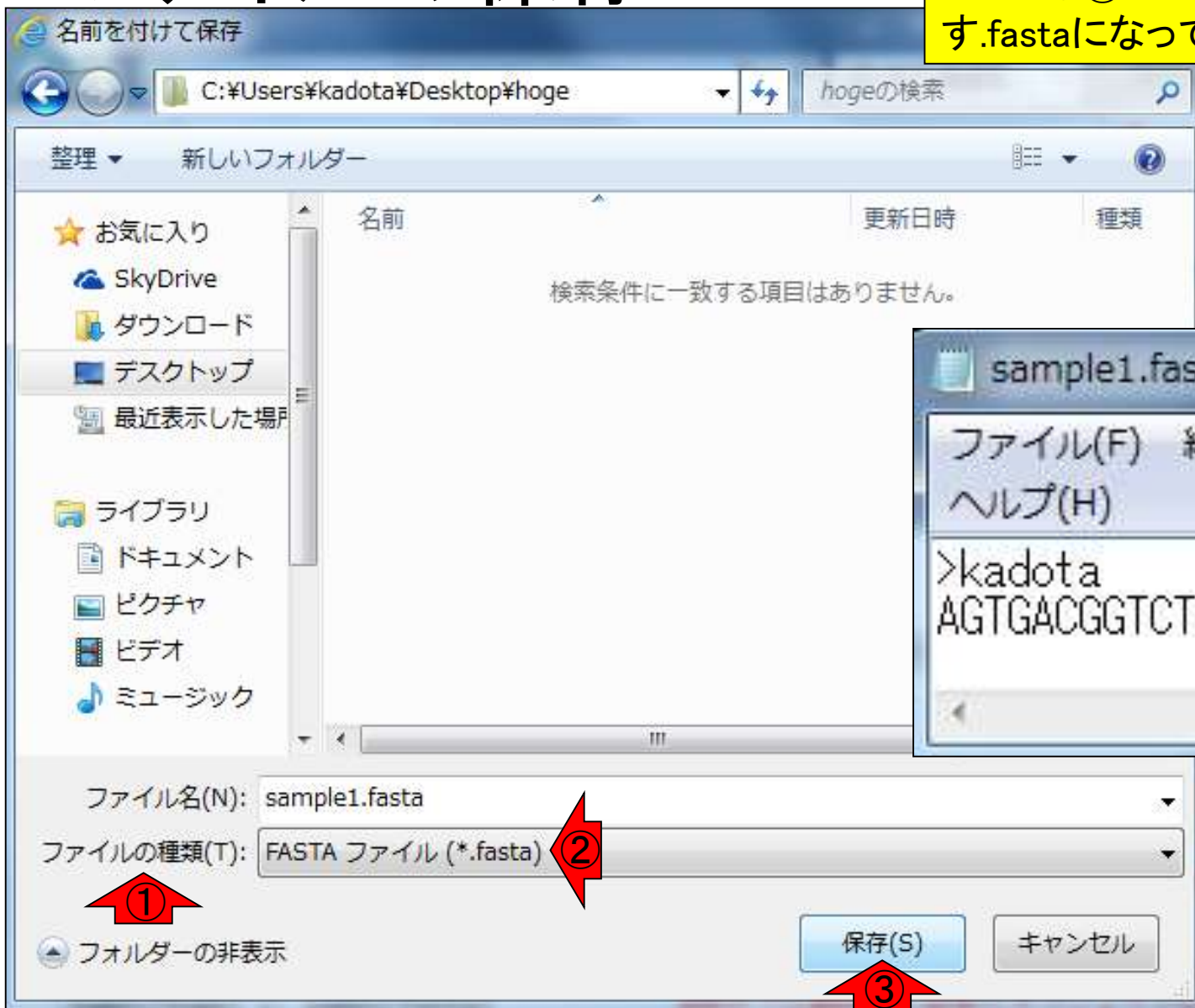
ファイルの種類(T): FASTA ファイル (*.fasta)

フォルダーの非表示

保存(S) **⑤** キャンセル

ファイルの保存

ときどき拡張子が*.txtなどと勝手に変わっていることがあるので①ファイルの種類欄に注意。ここでは②FASTA形式ファイルであることを示す.fastaになっていることを確認して③保存



作業ディレクトリの変更

R起動直後のデフォルトの作業ディレクトリは、
①ユーザ名 `kadota` のWindows環境では、「`C:/Users/kadota/Documents`」。その一方で、今解析したいディレクトリ(フォルダ)はデスクトップ上にある `hoge` なので、作業ディレクトリをそこに変更する必要がある。
②「`getwd()`」は、現在の作業ディレクトリを表示させるコマンド

R Console

```
R version 3.1.3 (2015-03-09) -- "Smooth Sidewalk"  
Copyright (C) 2015 The R Foundation for Statistical Computing  
Platform: x86_64-w64-mingw32/x64 (64-bit)
```

```
R は、自由なソフトウェアであり、「完全に無保証」です$  
一定の条件に従えば、自由にこれを再配布することができ$  
配布条件の詳細に関しては、'license()' あるいは 'lic$
```

```
R は多くの貢献者による共同プロジェクトです。  
詳しくは 'contributors()' と入力してください。  
また、R や R のパッケージを出版物で引用する際の形式$  
'citation()' と入力してください。
```

```
'demo()' と入力すればデモをみることができます。  
'help()' とすればオンラインヘルプが出ます。  
'help.start()' で HTML ブラウザによるヘルプがみられ$  
'q()' と入力すれば R を終了します。
```

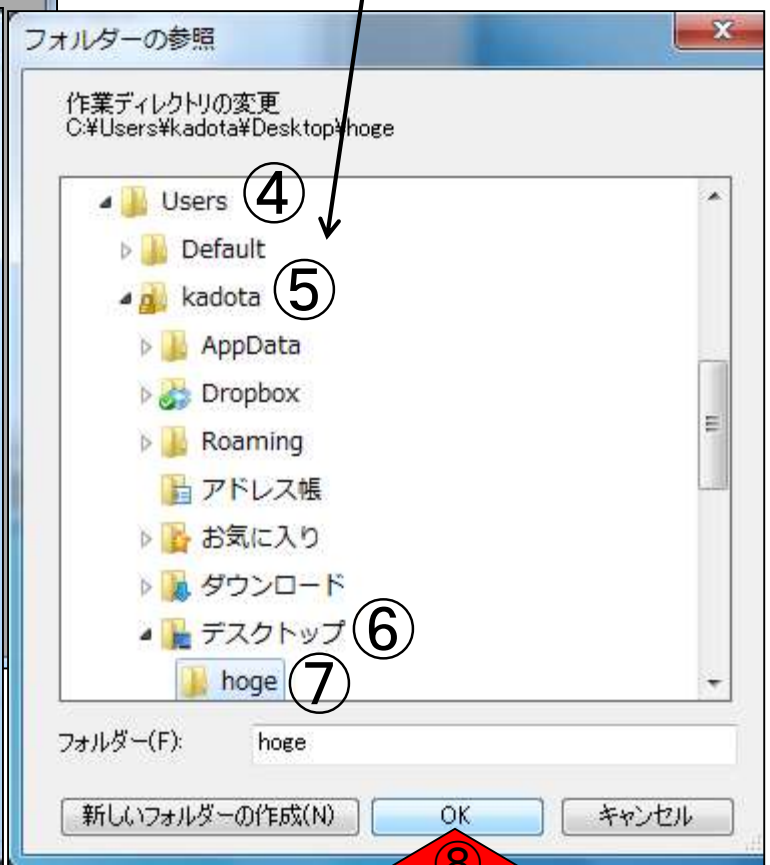
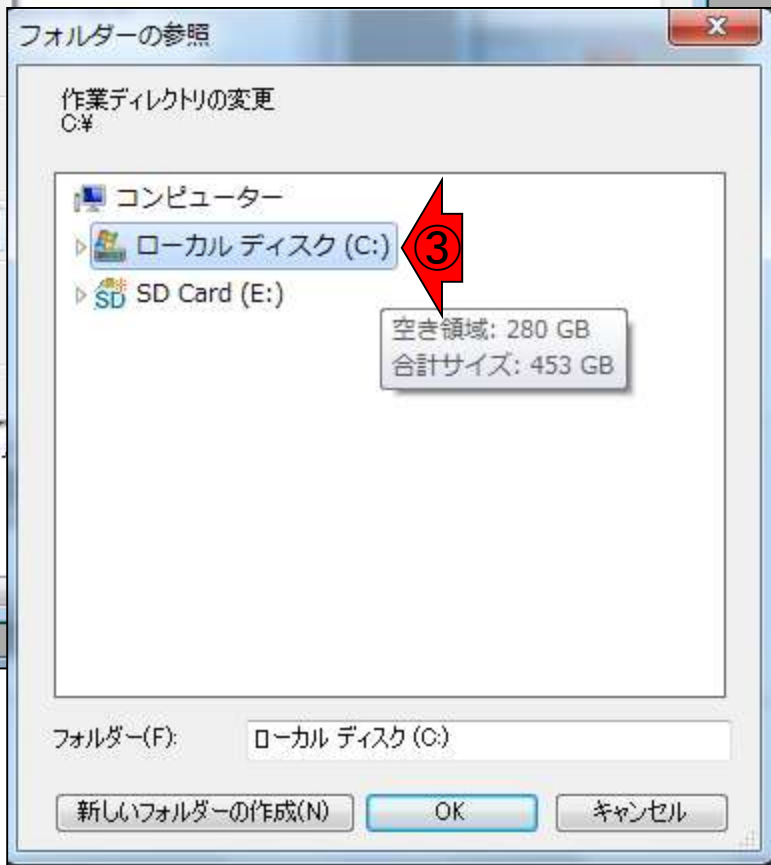
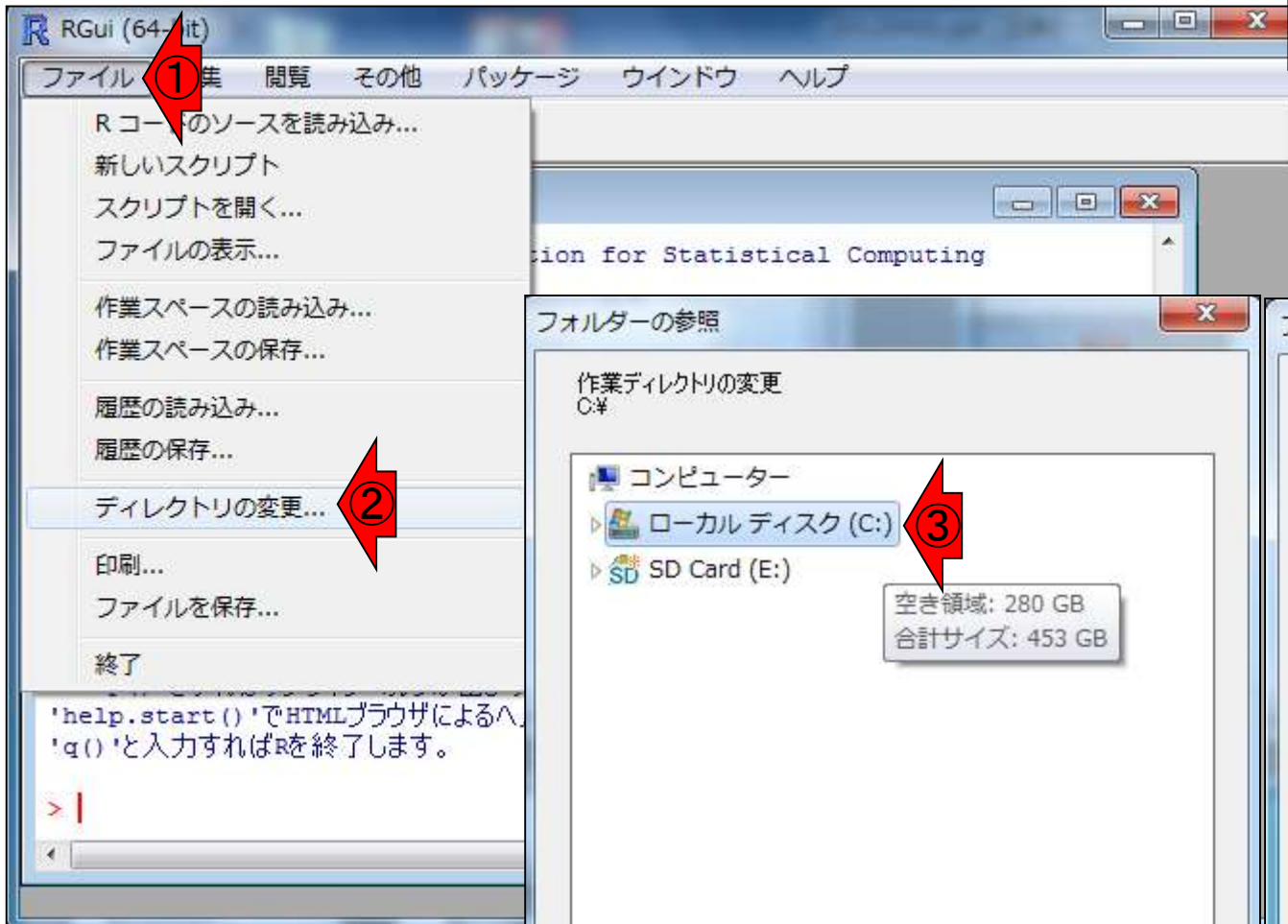
```
> getwd()  
[1] "C:/Users/kadota/Documents"  
> |
```

②

①

作業ディレクトリの変更

①ファイル、②ディレクトリの変更。
③「Windows(C:)」となっている場合もあるが、気にしない。⑤ヒトによって異なり、貸与PCの場合はiu



getwd()と打ち込んで確認

当たり前ですが、解析したいディレクトリ(またはフォルダ)を正しく指定できていなければエラーに遭遇します。また、解析したいファイルが存在しない状態でもエラーが出ます

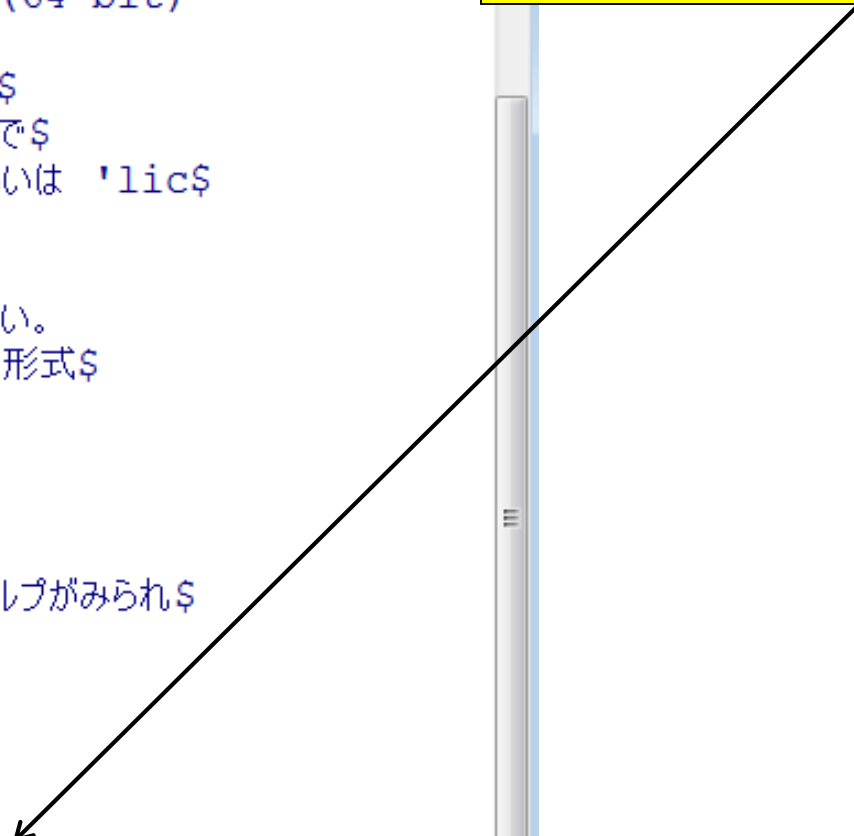
```
R Console
Platform: x86_64-w64-mingw32/x64 (64-bit)

R は、自由なソフトウェアであり、「完全に無保証」です$
一定の条件に従えば、自由にこれを再配布することがで$
配布条件の詳細に関しては、'license()'あるいは 'lic$

R は多くの貢献者による共同プロジェクトです。
詳しくは 'contributors()'と入力してください。
また、R や R のパッケージを出版物で引用する際の形式$
'citation()'と入力してください。

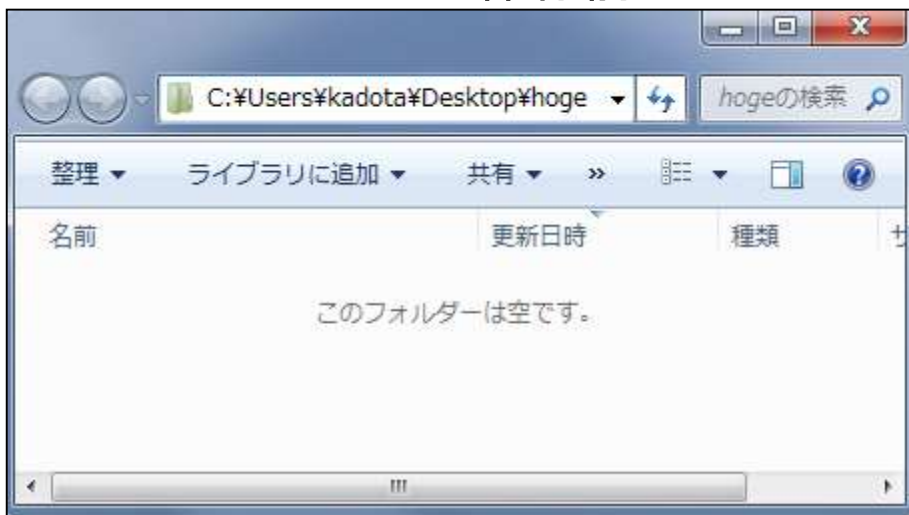
'demo()'と入力すればデモをみることができます。
'help()'とすればオンラインヘルプが出ます。
'help.start()'で HTML ブラウザによるヘルプがみられ$
'q()'と入力すれば R を終了します。

> getwd()
[1] "C:/Users/kadota/Documents"
> getwd()
[1] "C:/Users/kadota/Desktop/hoge"
> |
```

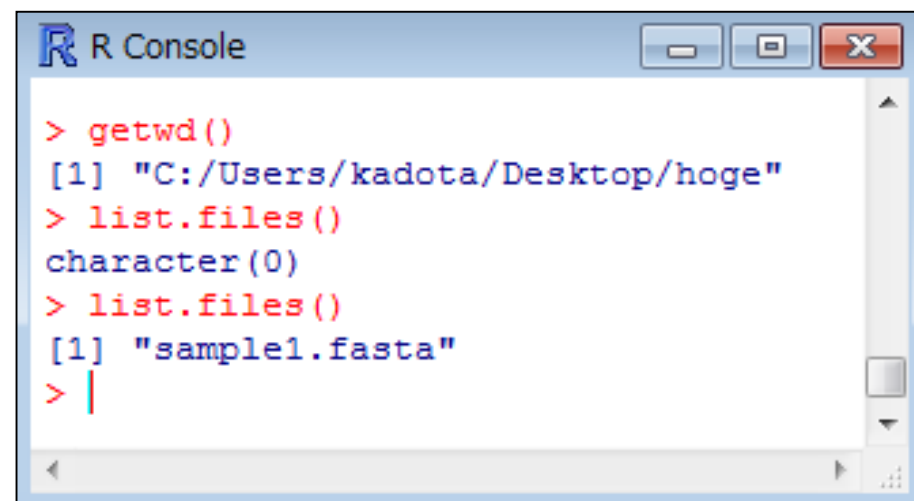
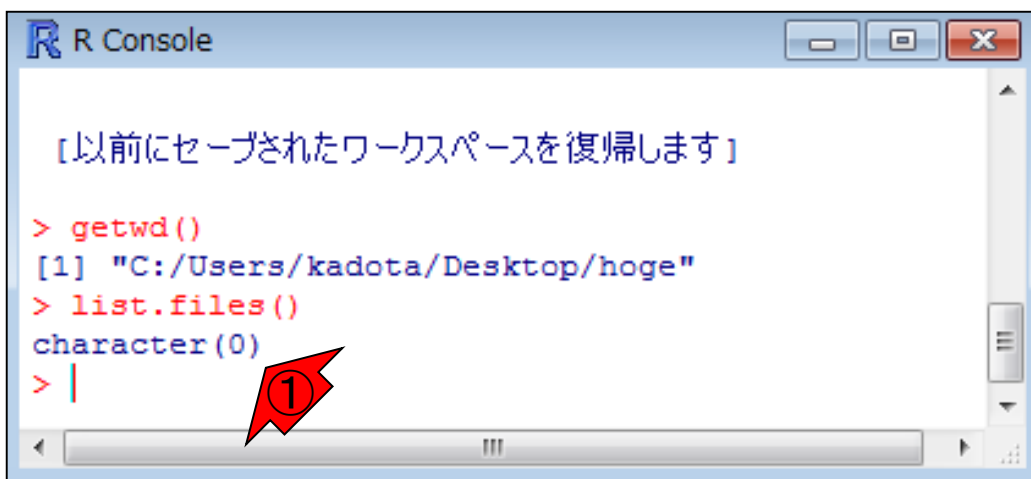
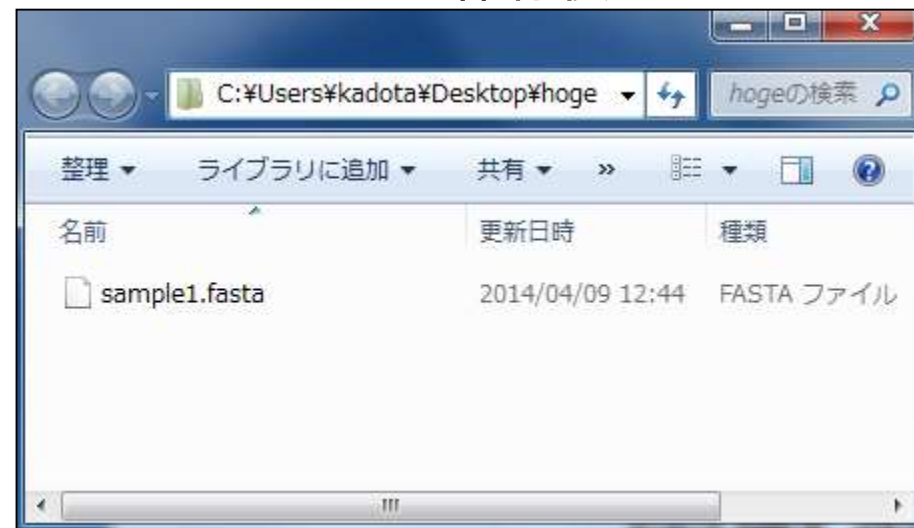


実際のhogeフォルダとR操作画面の関係

ファイル保存前



ファイル保存後



基本はコピー

イントロ | 一般 | 翻訳配列(translate)を取得 NEW

塩基配列を読み込んでアミノ酸配列に翻訳するやり方を示します。

「ファイル」-「ディレクトリの変更」で解析したいファイルを置いてあるディレクトリに移動し以下をコピー。

1. FASTA形式ファイル(sample1.fasta)の場合:

multi-FASTAではない single-FASTA形式ファイルです。

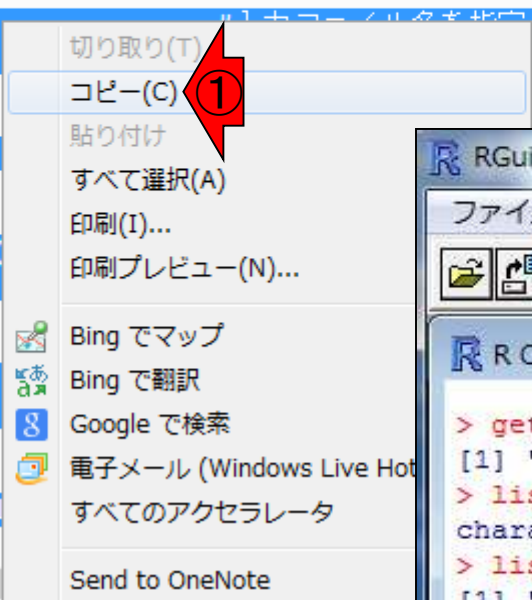
```
in_f <- "sample1.fasta"
out_f <- "hoge1.fasta"
```

```
#必要なパッケージをロード
library(Biostrings)
```

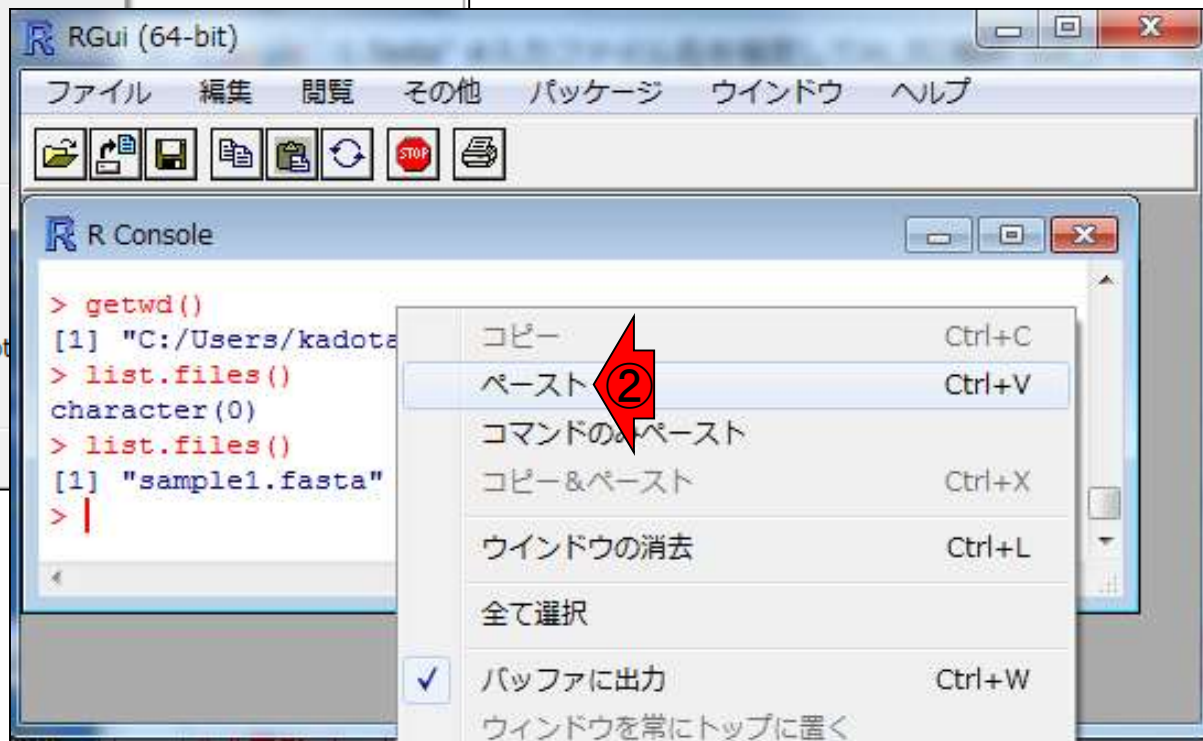
```
#入力ファイルの読み込み
fasta <- readDNASTringSet(
  fasta
```

```
#本番
fasta <- translate(fasta)
fasta
```

```
#ファイルに保存
writeXStringSet(fasta, fil
```



①一連のコマンド群をコピーして②R Console画面上でペースト。ブラウザがInternet Explorerの場合は、CTRLとALTキーを押しながらコードの枠内で左クリックすると、全選択できます。トリプルクリックでもよい。全選択の場合はできるかぎりこのやり方にしましょう



基本はコピペ

イントロ | 一般 | 翻訳配列(translate)を取得 NE

塩基配列を読み込んでアミノ酸配列に翻訳するやり方を示します。
「ファイル」-「ディレクトリの変更」で解析したいファイルを置いてあるディ

1. FASTA形式ファイル(sample1.fasta)の場合:

multi-FASTAではない single-FASTA形式ファイルです。

```

in_f <- "sample1.fasta"      #入力ファイル名
out_f <- "hogel.fasta"      #出力ファイル名

#必要なパッケージをロード
library(Biostrings)        #パッケージの読み込み

#入力ファイルの読み込み
fasta <- readDNASTringSet(in_f, format="fasta")#in_f
fasta                                #確認してるだけ

#本番
fasta <- translate(fasta)    #アミノ酸配列に
fasta                        #確認してるだけ

#ファイルに保存
writeXStringSet(fasta, file=out_f, format="fasta", w

```

エラーなく実行できた場合の全貌

```

> in_f <- "sample1.fasta"
> out_f <- "hogel.fasta"      #出力ファイル名を指定してout_fに格納
>
> #必要なパッケージをロード
> library(Biostrings)        #パッケージの読み込み
要求されたパッケージ BiocGenerics をロード中です
要求されたパッケージ parallel をロード中です

次のパッケージを付け加えます: 'BiocGenerics'

The following objects are masked from 'package:parallel':

clusterApply, clusterApplyLB, clusterCall, clusterEvalQ,
clusterExport, clusterMap, parApply, parCapply, parLapply,
parLapplyLB, parRapply, parSapply, parSapplyLB

The following object is masked from 'package:stats':

xtabs

The following objects are masked from 'package:base':

anyDuplicated, append, as.data.frame, as.vector, cbind,
colnames, do.call, duplicated, eval, evalq, Filter, Find, get,
intersect, is.unsorted, lapply, Map, mapply, match, mget, order,
paste, pmax, pmax.int, pmin, pmin.int, Position, rank, rbind,
Reduce, rep.int, rownames, sapply, setdiff, sort, table, tapply,
union, unique, unlist

要求されたパッケージ IRanges をロード中です
要求されたパッケージ XVector をロード中です
>
> #入力ファイルの読み込み
> fasta <- readDNASTringSet(in_f, format="fasta")#in_fで指定したファイルの読$
> fasta                                #確認してるだけです
  A DNASTringSet instance of length 1
    width seq                      names
[1]    12 AGTGACGGTCTT             kadota
>
> #本番
> fasta <- translate(fasta)            #アミノ酸配列に翻訳した結果をfasta$
> fasta                                #確認してるだけです
  A AAStringSet instance of length 1
    width seq                      names
[1]     4 SDGL                     kadota
>
> #ファイルに保存
> writeXStringSet(fasta, file=out_f, format="fasta", width=50)#fastaの中身を$
> |

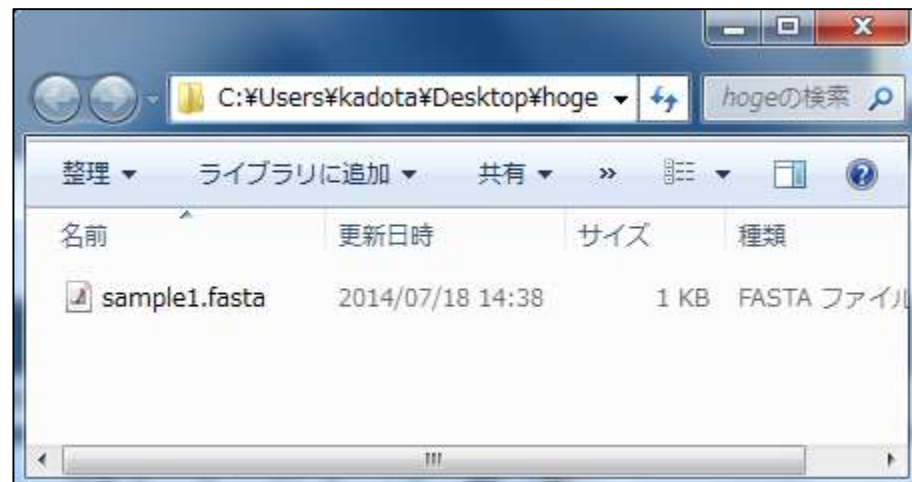
```

実行結果

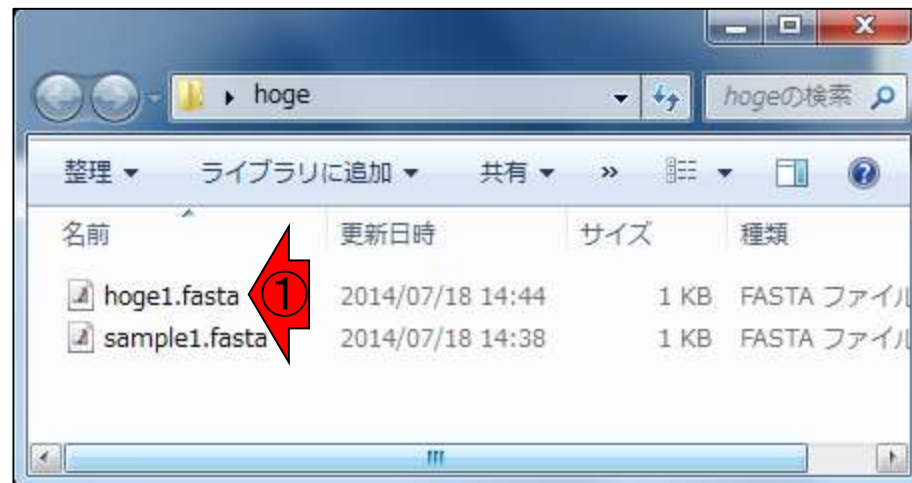
①出力ファイル名として指定したhoge1.fastaが生成されていることが分かります

```
R Console
> fasta <- readDNAStringSet(in_f, format="fasta")#$
> fasta #確認して$
A DNAStringSet instance of length 1
  width seq          names
[1]    12 AGTGACGGTCTT kadota
>
> #本番
> fasta <- translate(fasta) #アミノ酸$
> fasta #確認して$
A AAStringSet instance of length 1
  width seq          names
[1]     4 SDGL        kadota
>
> #ファイルに保存
> writeXStringSet(fasta, file=out_f, format="fasta$
> |
```

実行前のhogeフォルダ



実行後のhogeフォルダ



入出力の関係

イントロ | 一般 | 翻訳配列(translate)を取得 **NEW**

塩基配列を読み込んでアミノ酸配列に翻訳するやり方を示します。

「ファイル」-「ディレクトリの変更」で解析したいファイルを置いてあるディレクトリに移動し以下をコピペ。

1. FASTA形式ファイル(sample1.fasta)の場合:

multi-FASTAではないsingle-FASTA形式ファイルです。

```

in_f <- "sample1.fasta"      #入力ファイル名を指定してin_fに格納
out_f <- "hoge1.fasta"      #出力ファイル名を指定してout_fに格納

#必要なパッケージをロード
library(Biostrings)        #パッケージの読み込み

#入力ファイルの読み込み
fasta <- readDNAStringSet(in_f, format="fasta")#in_fで指定したファイルの読み込み
fasta                                     #確認してるだけです

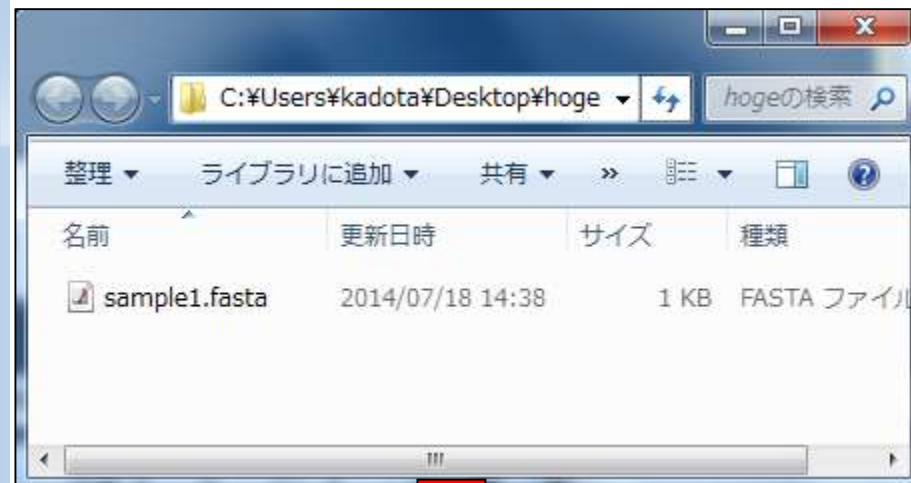
#本番
fasta <- translate(fasta)    #アミノ酸配列に翻訳した結果をfastaに格納
fasta                         #確認してるだけです

#ファイルに保存
writeXStringSet(fasta, file=out_f, format="fasta", width=50)#fastaの中身を指定したフ
    
```

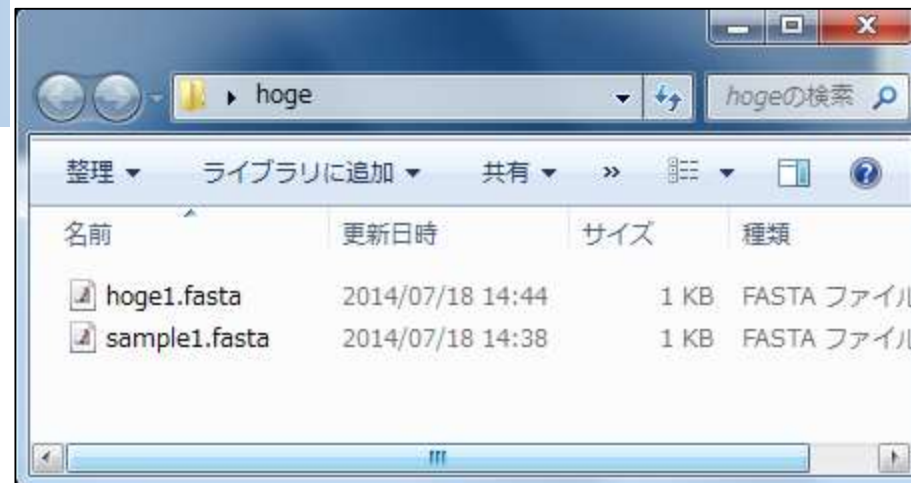
実行結果

①「list.files()で表示される結果」と②「実行後のhogeフォルダの中身」は当然同じ

実行前のhogeフォルダ



実行後のhogeフォルダ



```
R Console
> fasta <- readDNAStringSet(in_f, format="fasta")#$
> fasta                                     #確認して$
  A DNAStringSet instance of length 1
  width seq                                 $
[1]    12 AGTGACGGTCTT                       $
>
> #本番
> fasta <- translate(fasta)                 #アミノ酸$
> fasta                                     #確認して$
  A AAStringSet instance of length 1
  width seq                                 $
[1]     4 SDGL                               $
>
> #ファイルに保存
> writeXString(fasta, file=out_f, format="fasta")
> list.files()
[1] "hoge1.fasta"  "sample1.fasta"
> |
```



入力ファイル中の塩基配列は、3の倍数の12塩基長、ACGTのみからなるので何のエラーもない

実行結果

入力: 塩基配列ファイル (sample1.fasta)

```
>kadota
AGTGACGGTCTT
```

出力: アミノ酸配列ファイル (hoge1.fasta)

```
>kadota
SDGL
```



コドン表

<http://ja.wikipedia.org/wiki/%E3%82%B3%E3%83%89%E3%83%B3>

表1. 64コドンと各々に対応するアミノ酸を示したもの。mRNAの方向は5'から3'である。

		2nd base			
		U	C	A	G
1st base	U	UUU (Phe/F)フェニルアラニン	UCU (Ser/S)セリン	UAU (Tyr/Y)チロシン	UGU (Cys/C)システイン
		UUC (Phe/F)フェニルアラニン	UCC (Ser/S)セリン	UAC (Tyr/Y)チロシン	UGC (Cys/C)システイン
		UUA (Leu/L)ロイシン	UCA (Ser/S)セリン	UAA Ochre (終止)	UGA Opal (終止)
		UUG (Leu/L)ロイシン	UCG (Ser/S)セリン	UAG Amber (終止)	UGG (Trp/W)トリプトファン
	C	CUU (Leu/L)ロイシン	CCU (Pro/P)プロリン	CAU (His/H)ヒスチジン	CGU (Arg/R)アルギニン
		CUC (Leu/L)ロイシン	CCC (Pro/P)プロリン	CAC (His/H)ヒスチジン	CGC (Arg/R)アルギニン
		CUA (Leu/L)ロイシン	CCA (Pro/P)プロリン	CAA (Gln/Q)グルタミン	CGA (Arg/R)アルギニン
		CUG (Leu/L)ロイシン	CCG (Pro/P)プロリン	CAG (Gln/Q)グルタミン	CGG (Arg/R)アルギニン
	A	AUU (Ile/I)イソロイシン	ACU (Thr/T)スレオニン	AAU (Asn/N)アスパラギン	AGU (Ser/S)セリン
		AUC (Ile/I)イソロイシン	ACC (Thr/T)スレオニン	AAC (Asn/N)アスパラギン	AGC (Ser/S)セリン
		AUA (Ile/I)イソロイシン, (開始)	ACA (Thr/T)スレオニン	AAA (Lys/K)リシン	AGA (Arg/R)アルギニン
		AUG (Met/M)メチオニン, (開始) ^[3]	ACG (Thr/T)スレオニン	AAG (Lys/K)リシン	AGG (Arg/R)アルギニン
G	GUU (Val/V)バリン	GCU (Ala/A)アラニン	GAU (Asp/D)アスパラギン酸	GGU (Gly/G)グリシン	
	GUC (Val/V)バリン	GCC (Ala/A)アラニン	GAC (Asp/D)アスパラギン酸	GGC (Gly/G)グリシン	
	GUA (Val/V)バリン	GCA (Ala/A)アラニン	GAA (Glu/E)グルタミン酸	GGA (Gly/G)グリシン	
	GUG (Val/V)バリン, (開始)	GCG (Ala/A)アラニン	GAG (Glu/E)グルタミン酸	GGG (Gly/G)グリシン	

パッケージインストール確認

キーボードの上矢印キーを1回押すと直前に打ち込んだコマンドが表示される。もう一度リターンキーを押して実行すると、何のメッセージも表示されなくなる。これもエラーが出ていないのでOK

```
R Console
> library(Biostrings)
要求されたパッケージ BiocGenerics をロード中です
要求されたパッケージ parallel をロード中です

次のパッケージを付け加えます: 'BiocGenerics'

The following objects are masked from 'package:parallel':

  clusterApply, clusterApplyLB, clusterCall, clusterEvalQ,
  clusterExport, clusterMap, parApply, parCapply, parLapply,
  parLapplyLB, parRapply, parSapply, parSapplyLB

The following object is masked from 'package:stats':

  xtabs

The following objects are masked from 'package:base':

  anyDuplicated, append, as.data.frame, as.vector, cbind,
  colnames, do.call, duplicated, eval, evalq, Filter, Find, get,
  intersect, is.unsorted, lapply, Map, mapply, match, mget, order,
  paste, pmax, pmax.int, pmin, pmin.int, Position, rank, rbind,
  Reduce, rep.int, rownames, sapply, setdiff, sort, table, tapply,
  union, unique, unlist, unsplit

要求されたパッケージ S4Vectors をロード中です
要求されたパッケージ stats4 をロード中です
要求されたパッケージ IRanges をロード中です
要求されたパッケージ XVector をロード中です
> library(Biostrings)
> |
```

パッケージインストール確認

キーボードの上矢印キーなどを利用して、次にShortReadパッケージの確認を行う。エラーメッセージが出ていないことがわかる。

R Console

```
The following object is masked from 'package:stats':
```

```
xtabs
```

```
The following objects are masked from 'package:base':
```

```
anyDuplicated, append, as.data.frame, as.vector, cbind,
colnames, do.call, duplicated, eval, evalq, Filter, Find, get,
intersect, is.unsorted, lapply, Map, mapply, match, mget, order,
paste, pmax, pmax.int, pmin, pmin.int, Position, rank, rbind,
Reduce, rep.int, rownames, sapply, setdiff, sort, table, tapply,
union, unique, unlist, unsplit
```

```
要求されたパッケージ S4Vectors をロード中です
```

```
要求されたパッケージ stats4 をロード中です
```

```
要求されたパッケージ IRanges をロード中です
```

```
要求されたパッケージ XVector をロード中です
```

```
> library(Biostrings)
```

```
> library(ShortRead)
```

```
要求されたパッケージ BiocParallel をロード中です
```

```
要求されたパッケージ Rsamtools をロード中です
```

```
要求されたパッケージ GenomeInfoDb をロード中です
```

```
要求されたパッケージ GenomicRanges をロード中です
```

```
要求されたパッケージ GenomicAlignments をロード中です
```

```
> library(ShortRead)
```

```
> |
```


パッケージインストール確認

```
R Console  
  
intersect, is.unsorted, lapply, Map, mapply, match, mget, order,  
paste, pmax, pmax.int, pmin, pmin.int, Position, rank, rbind,  
Reduce, rep.int, rownames, sapply, setdiff, sort, table, tapply,  
union, unique, unlist, unsplit  
  
要求されたパッケージ S4Vectors をロード中です  
要求されたパッケージ stats4 をロード中です  
要求されたパッケージ IRanges をロード中です  
要求されたパッケージ XVector をロード中です  
> library(Biostrings)  
> library(ShortRead)  
要求されたパッケージ BiocParallel をロード中です  
要求されたパッケージ Rsamtools をロード中です  
要求されたパッケージ GenomeInfoDb をロード中です  
要求されたパッケージ GenomicRanges をロード中です  
要求されたパッケージ GenomicAlignments をロード中です  
> library(ShortRead)  
> library(Shortread)  
以下にエラー library(Shortread) :  
  'Shortread' という名前のパッケージはありません  
> library(shortread)  
以下にエラー library(shortread) :  
  'shortread' という名前のパッケージはありません  
> library(shortRead)  
以下にエラー library(shortRead) :  
  'shortRead' という名前のパッケージはありません  
> |
```



①Biostringsというパッケージは、②この部分でロードして使っていますが...同じ結果が得られる別のやり方も存在します

おまけ1: Biostrings

- ・ [イントロ](#) | [一般](#) | [任意の位置の塩基を置換](#) (last modified 2013/09/12)
- ・ [イントロ](#) | [一般](#) | [指定した範囲の配列を取得](#) (last modified 2015/04/06)
- ・ [イントロ](#) | [一般](#) | [指定したID\(染色体やdescription\)の配列を取得](#) (last modified 2014/03/10)
- ・ [イントロ](#) | [一般](#) | [翻訳配列\(translate\)を取得\(基礎\)](#) | [Biostrings](#) ① (last modified 2015/09/12)
- ・ [イントロ](#) | [一般](#) | [翻訳配列\(translate\)を取得\(応用\)](#) | [seqinr\(Chambers 2005\)](#) (last modified 2015/03/09)
- ・ [イントロ](#) | [一般](#) | [相補鎖\(complement\)を取得](#) (last modified 2013/06/14)

イントロ | 一般 | 翻訳配列(translate)を取得(基礎) | Biostrings NEW

Biostringsパッケージを用いて塩基配列を読み込んでアミノ酸配列に翻訳するやり方を示します。翻訳のための**遺伝コード(genetic code)**は、Standard Genetic Codeだそうです。もちろん生物種!!によって多少違い(variants)があるようで、"Standard", "SGC0", "Vertebrate Mitochondrial", "SGC1"などいろいろ選べるようです。「ファイル」-「ディレクトリの変更」で解析したいファイルを置いてあるディレクトリに移動し以下をコピー。

1. FASTA形式ファイル(sample1.fasta)の場合:

multi-FASTAではないsingle-FASTA形式ファイルです。

```

in_f <- "sample1.fasta"      #入力ファイル名を指定してin_fに格納
out_f <- "hoge1.fasta"      #出力ファイル名を指定してout_fに格納

#必要なパッケージをロード
library(Biostrings)        #パッケージの読み込み

#入力ファイルの読み込み
fasta <- readDNAStringSet(in_f, format="fasta") #in_fで指定したファイルの読み込み
                                              #確認してるだけです

#本番
fasta <- translate(fasta)   #アミノ酸配列に翻訳した結果をfastaに格納
fasta                      #確認してるだけです

#ファイルに保存
writeXStringSet(fasta, file=out_f, format="fasta", width=50) #fastaの中身を指定したフ

```

- ・ [イントロ](#) | [一般](#) | [逆相補鎖\(reverse complement\)](#) (last modified 2013/06/14)
- ・ [イントロ](#) | [一般](#) | [逆鎖\(reverse\)を取得](#) (last modified 2013/06/14)
- ・ [イントロ](#) | [一般](#) | [k-mer解析 | k=1\(塩基ごとの出現\)](#) (last modified 2013/06/14)
- ・ [イントロ](#) | [一般](#) | [k-mer解析 | k=2\(2連続塩基の出現\)](#) (last modified 2013/06/14)
- ・ [イントロ](#) | [一般](#) | [k-mer解析 | k=3\(3連続塩基の出現\)](#) (last modified 2013/06/14)
- ・ [イントロ](#) | [一般](#) | [k-mer解析 | k=n\(n連続塩基の出現\)](#) (last modified 2013/06/14)
- ・ (削除予定) [イントロ](#) | [一般](#) | [2連続塩基の出現](#) (last modified 2013/06/14)
- ・ (削除予定) [イントロ](#) | [一般](#) | [3連続塩基の出現](#) (last modified 2013/06/14)
- ・ (削除予定) [イントロ](#) | [一般](#) | [任意の長さの連続塩基の出現](#) (last modified 2013/06/14)
- ・ [イントロ](#) | [一般](#) | [Tips](#) | [任意の拡張子でファイル名を指定](#) (last modified 2013/06/14)
- ・ [イントロ](#) | [一般](#) | [Tips](#) | [拡張子は同じで任意の位置](#) (last modified 2013/06/14)

これです。①seqinrというパッケージを、
②この部分でロードして使っています

おまけ1: seqinr

- ・ イントロ | 一般 | [任意の位置の塩基を置換](#) (last modified 2013/09/12)
- ・ イントロ | 一般 | [指定した範囲の配列を取得](#) (last modified 2015/04/06)
- ・ イントロ | 一般 | [指定したID\(染色体やdescription\)の配列を取得](#) (last modified 2014/03/10)
- ・ イントロ | 一般 | 翻訳配列(translate)を取得(基礎) | [Biostrings](#) (last modified 2015/09/12)
- ・ イントロ | 一般 | 翻訳配列(translate)を取得(応用) | [seqinr\(Charif 2005\)](#) ① (last modified 2015/03/09)
- ・ イントロ | 一般 | [相補鎖\(complement\)を取得](#) (last modified 2013/06/14)

イントロ | 一般 | 翻訳配列(translate)を取得(応用) | [seqinr\(Charif_2005\)](#)

[seqinr](#)パッケージを用いて塩基配列を読み込んでアミノ酸配列に翻訳するやり方を示します。本気で翻訳配列を取得する場合にはこちらの利用をお勧めします。翻訳できないコドンにはアミノ酸X(不明なアミノ酸)に変換してくれたり、`translate`関数のオプションとして`ambiguous=T`とすると、翻訳できるものは可能な限り翻訳してくれます(高橋 広夫 氏提供情報)。`lapply`関数を用いるやり方(高橋 広夫 氏提供情報)と`sapply`関数を用いるやり方(甲斐 政親 氏提供情報)を示します。

「ファイル」-「ディレクトリの変更」で解析したいファイルを置いてあるディレクトリに移動し以下をコピー。

1. FASTA形式ファイル([sample1.fasta](#))の場合:

multi-FASTAではないsingle-FASTA形式ファイルです。

```

in_f <- "sample1.fasta"      #入力ファイル名を指定してin_fに格納
out_f <- "hoge1.fasta"      #出力ファイル名を指定してout_fに格納

#必要なパッケージをロード
library(seqinr)             #パッケージの読み込み

#入力ファイルの読み込み
hoge <- read.fasta(in_f, seqtype="DNA") #in_fで指定したファイルの読み込み
hoge                                #確認してるだけです

#本番
hoge <- lapply(hoge, function(x){ #アミノ酸配列に翻訳
  translate(x, ambiguous=T)      #アミノ酸配列に翻訳
})                               #アミノ酸配列に翻訳
hoge                             #確認してるだけです

#ファイルに保存
write.fasta(hoge, names=names(hoge), file.out=out_f, nbchar=50) #hogeの中身を指定したフ

```

おまけ2

①平成27年度NGSハンズオン講習会では、②Rのパッケージについてや、バージョンの違いに起因する問題など、より詳細な情報が7/29および7/30分の講義資料として提供されています

バイオインフォマティクス人材育成カリキュラム(次世代シーケンシング)

・バイオサイエンスデータベースセンター(NBDC)運営委員会人材育成分科会で、2014年3月に策定されたNGS用カリキュラムが存在。このカリキュラムをベースにして、平成26-29年度にハンズオン(ノートPCを用いた実習)を実施。おそらくアグリバイオ本体に次ぐ受講人数規模。

- ・ [NGS速習コース講習会](#) (平成26年度)
 - [NGS用カリキュラム](#)に沿った内容を東大農で2014年9月1日～10月10日に実施
 - 講習会映像は[統合TV](#)と [Youtube](#)から公開
 - [報告書](#)

①

- ・ [NGSハンズオン講習会](#) (平成27年度)
 - 東大農で2015年07月22日～08月06日 (A日程)および08月26日～09月05日 (B日程)に実施
 - 講習会映像は[統合TV](#)と [Youtube](#)から公開
 - [報告書](#)

- ・ [NGSハンズオン講習会](#) (平成28年度)
 - 東大農で2016年07月19日～08月04日に実施
 - 講習会映像は[統合TV](#)と [Youtube](#)から公開
 - [報告書](#)

- ・ [NGSハンズオン講習会](#) (平成29年度)
 - 東大農で2017年08月28日～09月01日に実施
 - 講習会映像は[統合TV](#)と [Youtube](#)から公開
 - [報告書](#)

実施日	実施時間	大項目	項目	レベル	習得技術	担当講師(敬称略)	講義資料・動画(統合TV)
7月22日 (水)	10:30-12:00	PC環境の構築	Bio-Linux8とRのインストール状況確認		・Linux導入 ・R導入 ・NGS解析に必要な環境構築技術	門田 幸二 (東京大学) 寺田 透 (東京大学)	事前予習資料 一巻 (PDF:52KB) 講義資料一巻 (PDF:108KB)
	13:15-14:45						
	15:00-16:30						
	16:45-18:15						
7月23日 (木)	10:30-12:00	UNIX/Linuxとスクリプト言語	Linux基礎	初級	UNIXの基礎の理解	門田 幸二 (東京大学)	講義資料一巻 (PDF:32KB) 統合TV
	13:15-14:45			中級			
	15:00-16:30						
	16:45-18:15						
7月24日 (金)	10:30-12:00		スクリプト言語	中級	シェルスクリプト	服部 恵美 (アメリカエフ)	講義資料 (PDF:1.8MB) 統合TV
	13:15-14:45						
	15:00-16:30						
	16:45-18:15						
7月27日 (月)	10:30-12:00			中級	Perl		講義資料 (PDF:1.9MB) 統合TV
	13:15-14:45						
	15:00-16:30						
	16:45-18:15						
7月28日 (火)	10:30-12:00			中級	Python		講義資料一巻 (PDF:52KB) 統合TV
	13:15-14:45						
	15:00-16:30						
	16:45-18:15						
7月29日 (水)	10:30-12:00	データ解析環境R	R基礎 1	初級	R言語の基礎 (インストールから利用まで)	門田 幸二 (東京大学)	講義資料一巻 (PDF:37KB) 統合TV
	13:15-14:45		R基礎 2	初級	・ファイルの読み込み ・行列演算の基本		
	15:00-16:30		R各種パッケージ	中級	パッケージのインストール法と代表的なパッケージの利用法		
	16:45-18:15						
7月30日 (木)	10:30-12:00		Bioconductorの利用法 1	中級	データの型やバージョンの違い		講義資料一巻 (PDF:53KB) 統合TV
	13:15-14:45						
	15:00-16:30		Bioconductorの利用法 2	中級	FASTA/FASTQファイルの各種解析		
	16:45-18:15						

②